Proceedings of

The International Conference on Artificial Intelligence Management and Trends

ISSN 3080-8103

ICAIMT

April 22-23, 2026

Abu Dhabi School of Management
Abu Dhabi
United Arab Emirates

Proceedings of

The International Conference on Artificial Intelligence Management and Trends

April 22-23, 2026

Abu Dhabi School of Management, Abu Dhabi, UAE

Organizing Committee:

Dr. Neda Abdelhamid (Chair)
Chair of MSBA Program
Associate Professor of Business Analytics

Dr. Constance Van Horne
Director of International Affairs
& Associate Professor of Management
Abu Dhabi School of Management

Editorial Board

The editorial board members for this conference are:

Preface

The *International Conference on Artificial Intelligence Management and Trends (ICAIMT)* was successfully held on April 22-23 2026, at the Abu Dhabi School of Management. This conference brought together industry experts, researchers, and practitioners to explore emerging trends in artificial intelligence, including automation, edge computing, and ethical AI, and to demonstrate how these advancements can address real-world challenges.

ICAIMT highlighted innovative tools and methodologies that empower businesses to make data-driven decisions, enhance operational efficiency, and foster industry-wide innovation.

This dynamic event provided participants with valuable insights and practical applications, encouraging interactive learning and meaningful engagement. It also served as a platform for collaboration, opening doors to future research, partnerships, and projects in this rapidly evolving field.

Table of Contents

Title **Page**

Artificial Intelligence in Management and Digital sMarketing: Emerging Trends, Organisational Transformations, and Ethical Implications

M. Chouaib DAKOUAN
Professor in Management at Hassan II University. Faculty of Legal, Economic and Social Sciences Ain Sebaa. Morocco.
Chouaib.dakouan@gmail.com

Mrs. Oumaima LEBBAR
Ph.D candidate at ISTEC Business School Paris. France.
Oumaimalebbar1@gmail.com

Abstract— This paper explores the ways Artificial Intelligence is transforming both managerial practices and digital marketing strategies across contemporary organisations. Drawing on recent scholarly work, it adopts a conceptual perspective to outline several emerging trends shaped by AI, such as predictive data analysis, smart automation, personalised customer interactions, and decision-making grounded in extensive datasets. The study further examines the organisational consequences of integrating AI, including shifts in internal structures, evolving skill demands, and the need for strengthened governance. Early observations indicate that AI can boost performance levels and deepen user engagement, yet it also introduces ethical questions involving transparency, algorithmic bias, and the safeguarding of personal data. In response, the paper proposes a unified framework designed to guide the responsible and efficient use of AI within both management and marketing contexts.

Keywords— Artificial Intelligence; Management Innovation; Digital Marketing; Predictive Analytics; Marketing Automation; Organisational Transformation; Data-Driven Decision-Making; Customer Experience; Ethical AI; Algorithmic Governance; Personalisation; Automation.

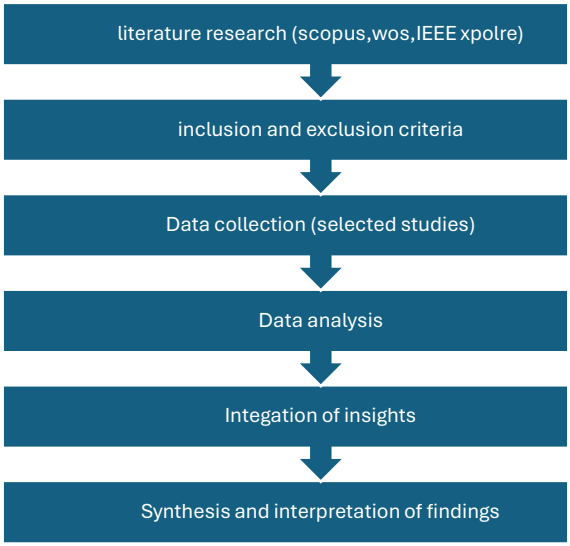
I. INTRODUCTION

The rapid evolution of Artificial Intelligence has significantly reshaped managerial frameworks as well as the digital marketing mechanisms used within both public and private organisations. AI tools now play a central role in guiding strategic choices, streamlining operational workflows, and redefining how organisations connect with and understand their audiences. Although digital transformation has been widely studied, notable gaps persist in research that examines the combined effects of AI on management and digital marketing[1]. Much of the existing literature tends to isolate these domains, analysing either managerial changes or marketing innovations separately, without fully recognising how closely intertwined they have become in data-centric organisational environments.[2]

This study seeks to address this limitation by presenting an integrated analytical view that captures the influence of AI on decision-making processes, organisational restructuring, and the enhancement of digital customer experiences[1]. It emphasizes key emerging trends such as predictive modelling, intelligent automation, and hyper-personalised marketing journeys[3], while also identifying the internal adjustments organisations must undertake—including new skill sets, adapted governance models, and cross-functional data practices. In addition, the paper underscores the importance of ethical vigilance, as AI-driven systems introduce concerns related to privacy protection, algorithmic fairness, and transparency in automated decisions.

By combining managerial and marketing perspectives within a single framework, this work provides managers, researchers, and digital strategists with a comprehensive foundation for understanding how AI can be leveraged responsibly and effectively. It also offers insights that may support future organisational policies, capability development, and the design of sustainable AI-driven strategies.[1]

II. METHODOLOGY



This study is grounded in a qualitative conceptual approach supported by an extensive review of recent scholarly work. Peer-reviewed publications appearing between 2018 and 2024 were gathered from Scopus, Web of Science, IEEE Xplore, and leading journals in the field of digital marketing. The selection process targeted research that specifically examines the role of Artificial Intelligence in managerial practices, marketing automation systems, customer analytics, and broader organisational transformations.

To interpret the collected material, the analysis relied on several complementary tools, including thematic content examination, the mapping of emerging trends, and the comparison of established managerial and marketing models. When relevant, additional insights were drawn from reports issued by international organisations and major technology firms, allowing the academic findings to be supported with perspectives from industry and policy contexts.

III. PRELIMINARY RESULTS AND INNOVATIVE FRAMEWORK

The preliminary results of the study point to four major developments shaping the integration of Artificial Intelligence within contemporary organisations.

First, managerial models are undergoing significant transformation as intelligent automation tools and predictive dashboards begin to influence strategic planning, performance monitoring, and the management of human resources. These technologies support more proactive, data-informed decision-making and introduce new forms of operational coordination.

Second, digital marketing functions are becoming markedly more sophisticated. AI applications strengthen audience segmentation, enable deeper personalisation, improve campaign optimisation, and support interactive tools such as chatbots. They also provide more detailed insights into customer journeys, allowing organisations to better anticipate needs and tailor engagement strategies.

Third, a culture centred on data is taking root across organisations. Effective governance practices and a solid understanding of AI concepts among employees are emerging as essential elements for ensuring the long-term, responsible use of these technologies. This shift requires systematic investment in skills, processes, and data infrastructure.

Fourth, the adoption of AI raises critical ethical and regulatory questions. Concerns surrounding algorithmic bias, the transparency of automated decisions, and the protection of personal data underscore the need for strong governance mechanisms capable of ensuring compliance and safeguarding stakeholder trust.

Building on these findings, the paper proposes an integrated AI–Management–Digital Marketing Framework. This model brings together strategic, operational, and ethical dimensions to guide organisations in deploying AI coherently and responsibly across both managerial and marketing domains.

IV. CONCLUSION

This study underscores the pivotal influence of Artificial Intelligence as a driving force in both managerial evolution and digital marketing innovation. The results highlight the necessity for organisations to adopt integrated strategies that bring together technological advancement, human capabilities, and sound governance practices. Such an approach is essential for ensuring that AI tools are deployed in ways that create value while maintaining organisational coherence and accountability.

Looking ahead, future research could include empirical testing of the proposed framework to assess its applicability in real-world settings. Comparative studies across different industries would also offer valuable insights into how sector-specific factors shape AI adoption. In addition, examining the diffusion of AI within emerging markets may reveal unique opportunities and constraints that are not yet fully captured in the literature. Strengthening ethical standards and expanding digital skill sets will remain critical priorities to guarantee the sustainable and trustworthy integration of AI across organisational functions.

REFERENCES

[1] Bouma, G., & Wessels, J. (2023). Artificial intelligence and organisational decision-making: Challenges and opportunities. Journal of Management Studies, 60(4), 845–864.

[2] Chaffey, D., & Ellis-Chadwick, F. (2022). *Digital Marketing* (8th ed.). Pearson.

[3] Huang, M.-H., & Rust, R. (2021). Artificial intelligence in service. *Journal of Service Research*, 24(1), 3–20.

[4] Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? *Business Horizons*, 62(1), 15–25.

[5] Wirtz, J., Patterson, P., Kunz, W., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2023). Brave new world: Service robots in marketing. *Journal of Marketing*, 87(2), 22–45.

Advanced Time Series Forecasting Models for Electricity Consumption Prediction in Abu Dhabi

Jaber Jemai
CIS division
Higher Colleges of Technoogy
Abu Dhabi, UAE
jjemai@hct.ac.ae

Alia Al Hammadi
CIS division
Higher Colleges of Technoogy
Abu Dhabi, UAE
H00492493@hct.ac.ae

Alya Al Zaabi
CIS division
Higher Colleges of Technoogy
Abu Dhabi, UAE
H00503082@hct.ac.ae

Rawdha Al Zaabi
CIS division
Higher Colleges of Technoogy
Abu Dhabi, UAE
H00492127@hct.ac.a

Abstract - Faced with the constant rise in energy demand and the urgent need to combat climate change, many governments have been driven to wisely consume available energy and explore new resources through innovative and sustainable solutions. Accurate forecasting of electricity demand is critical for operational planning and grid optimization. In this study, we aim to design and develop a comprehensive set of time series forecasting models to predict power consumption for TAQA, the main energy distributor in Abu Dhabi. Using historical load data, we will construct and compare four families of models: baseline econometric models (ARMA and ARIMA), a gradient-boosted tree ensemble (XGBoost), and a deep learning architecture (LSTM). In addition to forecasting performance, the models will undergo rigorous validation using a suite of time-series–specific diagnostic tests, including assessments of heteroscedasticity, autocorrelation, stationarity, and residual distribution properties. These diagnostics ensure the stability and statistical reliability of the models developed. The forecast accuracy of all models will be evaluated across multiple horizons using standard error metrics.

Keywords—Electricity Consumption Forecasting, Time Series Analysis, LSTM, XGBoost.

I. INTRODUCTION

Abu Dhabi is a fast-growing city that faces increasing energy demand driven by rapid urbanization, rising population, and industrial growth. This leads to higher costs of energy production and supply. The technical report of the Department of Energy (DoE) shows that the electricity demand in Abu Dhabi is continuously increasing to reach 13 GigaWatts (GW) in 2023, with a 9.1% increase from 2022 to 2023. TAQA Distribution, as the primary electricity and water distribution company in Abu Dhabi, plays a significant role in addressing these challenges. Such figures reveal a need for reliable and robust solutions to support TAQA in optimizing its electricity distribution grid.

Predicting electricity demand is one of the primary challenges faced by energy supply companies. If future demands are accurately approximated, then energy production factories and distribution grids will operate efficiently, peak demands will be known ahead of time, electricity shortages will be avoided, and customer satisfaction will be improved. To achieve this, TAQA aims to develop and deploy energy consumption forecasting solutions utilizing advanced econometric and AI-enabled models to support Abu Dhabi and the United Arab Emirates' vision for sustainability and its Net Zero 2050 goals. In this paper, we will use historical energy consumption data from TAQA to create short and long-term forecast. The dataset will include historical electricity usage of Abu Dhabi, Al Ain, and Al Dhafra regions. The goal of this paper is to test and compare advanced econometric and AI-enabled models to make the most accurate predictions. Similar studies have been conducted in various countries, including the United States of America [1], China [2], Brazil [3], and Mexico [4].

This project adds significant value to TAQA and other stakeholders by providing a practical solution for short and long-term energy demand forecasting. Through the use of machine learning techniques such as XGBoost and LSTM, the project aims to identify the most accurate predictive model. By analyzing historical energy consumption data alongside external factors like temperature and time patterns, the solution helps reduce uncertainty in demand management. For TAQA, this translates into better resource allocation, cost reduction, and more efficient energy planning. Ultimately, the project resolves the pressing problem of inefficiencies and rising costs in energy management by offering a scalable, intelligent solution.

II. RESEARCH METHODOLOGY

This study follows a structured methodological framework (Fig. 1) consisting of data acquisition, preprocessing, model development, diagnostic validation, and performance evaluation. The methodology ensures systematic comparison between all models to identify the most reliable forecasting approach for TAQA's electricity consumption.

Fig. 1. Research Methodology

III. LITERATURE REVIEW

The application of Artificial Intelligence (AI) and Machine Learning (ML) techniques have significantly transformed electricity consumption forecasting, addressing the limitations of traditional statistical methods that struggle to capture nonlinear and dynamic energy patterns [1]. Early models, such as ARIMA and regression analysis, provided foundational approaches, but proved inadequate when dealing with complex temporal dependencies and exogenous influences [3].

Consequently, AI-based models have emerged as more powerful alternatives capable of enhancing prediction accuracy. Machine learning approaches, including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and ensemble algorithms such as XGBoost and CatBoost have been effectively applied in diverse contexts. For instance, studies in the USA used Energy Information Administration (EIA) and Department of Energy (DOE) [5] datasets demonstrated that ML-based forecasting outperformed traditional methods in modeling nonlinear consumption behaviors, with LR and RF achieving notable accuracy gains. Similarly, CatBoost combined with hybrid optimization algorithms such as Particle Swarm Optimization (PSO) and Ant Lion Optimization (ALO) proved effective in long-term national-scale forecasting for sustainable energy planning in Mexico [6].

Deep learning techniques, particularly LSTM networks and their variants, have become dominant in short-term and medium-term load forecasting due to their ability to model sequential dependencies [7]. Comparative analyses show that Bidirectional LSTM (Bi-LSTM) models achieve superior performance relative to conventional LSTM, ARIMA, and SARIMA methods, particularly for nonlinear and seasonal datasets [8]. Recent advancements, such as the Improved LSTM (ILSTM) [9] integrated with IoT-based data acquisition systems, achieved outstanding results in residential energy forecasting within the United Arab Emirates, with MAE and RMSE of 0.185 and 0.276, respectively.

Hybrid AI architecture has further improved forecasting accuracy by integrating deep learning with optimization algorithms. The Optimized Deep Neural Network (ODNN) enhanced through a hybrid Genetic Algorithm–Particle Swarm Optimization (GA-PSO) achieved higher accuracy and lower error than conventional deep networks [10]. Similarly, Least Squares Support Vector Machine (LSSVM) tuned via the Gooseneck Barnacle Optimizer (GBO) [11] reached exceptional forecasting accuracy, outperforming several other bio-inspired optimizers such as PSO and Dragonfly Algorithm. Combining machine learning and fuzzy logic have demonstrated strong potential in managing uncertainty within microgrids. A study integrating BiLSTM forecasting with Mamdani fuzzy logic [12] achieved 93.2% classification accuracy for microgrid stability assessment. Hybrid optimization methods such as Echo State Networks (ESN) optimized by Differential Evolution (DE) [4] achieved very low forecasting errors using datasets from China and Taiwan. Comprehensive surveys of AI applications in energy systems report that hybrid models integrating deep learning with swarm intelligence or evolutionary optimization achieve the highest accuracy and reliability [1,13]. As reported in [2,14], these models not only enhance prediction accuracy but also support intelligent control and optimization in energy management, as seen in applications of AI to hybrid renewable systems and building-level forecasting.

IV. TAQA DATASET AND DATA PREPROCESSING

Time series data plays a central role in energy forecasting, as it captures changes in consumption over time with strong temporal dependencies. In the context of electrical grids and regional demand analysis, TAQA will provide a dataset reporting electricity consumption in the Emirate of Abu Dhabi. The dataset spans energy usage between 2019 and 2025 on hourly basis. Geographically, data will be covering Abu Dhabi City, Al Ain, and Al Dhafra regions at sub-station granularity level. Additionally, data on humidity, temperature, and precipitation will be collected from the National Center of Meteorology in the UAE.

Effective preprocessing is essential because real-world energy data often contains frequent irregularities, such as missing values, noise, and peaks, caused by unpredictable events. To clean TAQA's dataset, we expect to handle gaps in the time series when timestamps are missing. To reconstruct the timeline at uniform intervals, typical missing values imputation techniques will be used. It is also important to separate and process legitimate and erroneous extreme spikes. Furthermore, feature engineering of raw timestamps will be undertaken to extract time-based information such as hour-of-day, day-of-week, or season. These temporal attributes allow models to recognize recurring behaviors such as lag features, rolling windows, moving averages, and seasonal growth. This makes an early understanding of the critical characteristics of the time series. Properly addressing these and similar issues will transform raw observations into informative time-aware variables capable of revealing trends, recurring patterns, and long-term dependencies.

V. FORECASTING MODELS SELECTION

The selected models include classical econometric approaches, along with advanced machine learning and deep learning models. Together, these models provide a comprehensive forecasting framework combining traditional statistical principles with modern artificial intelligence methods.

A. Baseline forecasting models

Primarily, time-series forecasting relied on econometric models such as Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA). These models are grounded in statistical theory and remain widely used due to their interpretability and effectiveness in modeling linear temporal patterns. In the context of energy forecasting, econometric models help establish a baseline understanding of consumption trends by capturing dependencies within historical data. They are especially useful when consumption follows stable, predictable cycles. The ARMA model combines two components: an autoregressive term that expresses current values as a function of previous values, and a moving average term that incorporates past forecast errors. If energy consumption data exhibits non-stationarity due to seasonal effects, temperature variation, and long-term trends, then ARIMA models can be used.

B. XGBoost models

XGBoost is an advanced ensemble model based on gradient boosting principles by building trees sequentially, with each tree correcting the errors of the previous one. It incorporates strong regularization terms to prevent overfitting and includes features such as parallel computation, missing-value handling, and optimized memory usage. This makes XGBoost efficient, scalable, and highly accurate for forecasting tasks involving tabular datasets with many engineered features. XGBoost performs well in structured data environments because it can model high-level feature interactions, such as relationships between temperature, calendar effects, and regional consumption trends.

C. XGBoost models

Long short-term memory networks are a type of recurrent neural network designed specifically to model sequential and time-dependent data. LSTMs can automatically learn temporal dependencies by processing sequences such as the past 24 or 48 hours of consumption data. This makes LSTMs highly effective for modeling cyclic patterns and long-range temporal behavior. LSTMs excel in environments where energy consumption is influenced by strong daily and weekly cycles.

VI. CONCLUSION

This paper presented a data-driven methodological framework for forecasting TAQA's hourly electricity demand using representative statistical, machine learning, and deep learning models. Through a focused literature review and systematic model selection, ARMA, ARIMA, XGBoost, and LSTM were identified as suitable approaches for capturing both linear and nonlinear load dynamics. The proposed methodology emphasizes rigorous preprocessing and diagnostic validation, ensuring that each model meets key time-series reliability criteria. Overall, the framework provides a solid foundation for developing accurate and scalable forecasting solutions in energy analytics. Future work will apply this methodology to full empirical evaluation to TAQA dataset.

REFERENCES

[1.] R. Klyuev, I. Morgoev, A. Morgoeva, O. Gavrina, N. Martyushev, E. Efremenkov, and Q. Mengxu. Methods of forecasting electric energy consumption: A literature review. Energies, 15(23):8919, 2022.

[2] G. Hafeez, Kh. S. Alimgeer, Z. Wadud, Z. Shafiq, M. Ali Khan, I. Khan, F. Khan, and A. Derhab. A novel accurate and fast converging deep learning-based model for electrical energy consumption forecasting in a smart grid. Energies, 13(9):2244, 2020.

[3] J. de Assis Cabral, L. Legey, and M. de Freitas Cabral. Electricity consumption forecasting in brazil: A spatial econometrics approach. Energy, 126:124–131, 2017. \

[4] L. Wang, H. Hu, Xue-Yi Ai, and H. Liu. Effective electricity energy consumption forecasting using echo state network improved by differential evolution algorithm. Energy, 153:801–815, 2018.

[5] S. Hossain, M. Hasanuzzaman, M. Hossain, M. Amjad, M. Shovon, M. Hossain, and M. Rahman. Forecasting energy consumption trends with machine learning models for improved accuracy and resource management in the usa. Journal of Business and Management Studies, 7(1):200–217, 2025.

[6] H. Li and A. Kayae. Predicting energy consumption in mexico: Integrating environmental, economic, and energy data with machine learning techniques for sustainable development. Energy, 324:135992, 2025.

[7] M. W. Hasan. Design of an iot model for forecasting energy consumption of residential buildings based on improved long short-term memory (lstm). Measurement: Energy, 5:100033, 2025.

[8] H. Alizadegan, B. Rashidi Malki, A. Radmehr, H. Karimi, and M. Ilani. Comparative study of long short-term memory (lstm), bidirectional lstm, and traditional machine learning approaches for energy consumption prediction. Energy Exploration & Exploitation, 43(1):281–301, 2025.

[9] E. Uzun, M. Karabinao˘glu, and M. Ayaz. Lstm-based estimation of energy consumption in energy-intensive facilities. Engineering Research Express, 7(1):015420, 2025.

[10] E. Hosseini, B. Saeedpour, M. Banaei, and R. Ebrahimy. Optimized deep neural network architectures for energy consumption and pv production forecasting. Energy Strategy Reviews, 59:101704, 2025.

[11] M. Ahmed, M. Sulaiman, M. Hassan, M. Rahaman, and M. Amin. Daily allocation of energy consumption forecasting of a power distribution company using optimized least squares support vector machine. Results in Control and Optimization, 18:100518, 2025.

[12] Y. Kholiavka and Y. Parfenenko. Intelligent energy consumption forecasting and microgrid state assessmen

A Clinical AI Framework for Improving Physiotherapy Referral Decisions in Outpatient Care

Hajer Sayyar AlHosani
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
adsm-215601@adsm.ac.ae

Ishtiaq Rasool Khan
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
i.khan@adsm.ac.ae
ORCID: 0000-0002-3887-9052

Abstract— Artificial intelligence (AI) is increasingly applied to enhance clinical decision-making and reduce unwarranted variability in healthcare delivery. In outpatient rehabilitation settings, physiotherapy referral decisions are frequently based on individual clinician judgment, which may lead to delayed intervention, inconsistent access to rehabilitation services, and inefficiencies in patient pathways. This paper presents a short conceptual framework proposing an AI-enabled clinical decision support system (CDSS) designed to support physiotherapy referral decisions using structured electronic medical record (EMR) data.

The proposed framework integrates supervised machine learning models, explainable artificial intelligence mechanisms, enterprise-level deployment architecture, and a governance-aligned implementation strategy. Unlike a full experimental study, this work focuses on the conceptual design of an operational AI system, including the clinical workflow, methodological justification, and validation strategy required for real-world deployment. The framework aims to enhance referral consistency, improve operational efficiency, and support equitable access to rehabilitation services while preserving physician oversight and clinical governance. Future work will involve pilot implementation and empirical validation within outpatient rehabilitation settings.

Keywords— Artificial Intelligence, Clinical Decision Support Systems, Physiotherapy Referral, Machine Learning, Healthcare Governance

I. INTRODUCTION

Artificial intelligence (AI) is transforming healthcare delivery by enabling data-driven decision support and improving the consistency of clinical practice. The increasing adoption of electronic medical records (EMRs) has created large volumes of structured clinical data that can be analyzed using machine learning (ML) methods to support clinical decision-making processes. AI-enabled clinical decision support systems (CDSS) are increasingly used to assist clinicians in risk prediction, triage prioritization, and treatment planning.

In outpatient rehabilitation services, physiotherapy referral decisions play a critical role in determining patient recovery trajectories and healthcare resource utilization. Early referral to physiotherapy can significantly improve functional outcomes, reduce chronic disability risk, and enhance patient satisfaction. However, referral decisions are often influenced by individual clinician experience, workload pressures, and variability in clinical interpretation. As a result, similar patients may receive different referral decisions across providers or clinical settings.

The widespread adoption of EMR systems provides an opportunity to support referral decisions using predictive analytics. Machine learning models can analyze multidimensional patient data and identify patterns associated with successful physiotherapy intervention. By integrating these predictive insights within a CDSS framework, clinicians may receive structured decision support that complements clinical judgment.

This paper presents a short conceptual framework for an AI-enabled CDSS designed to support physiotherapy referral decisions in outpatient care. Rather than presenting a full experimental implementation, the study focuses on system design, methodological justification, and practical considerations for real-world deployment. The framework integrates machine learning models, explainable AI mechanisms, enterprise deployment architecture, and governance-aligned implementation strategies to facilitate responsible AI adoption in clinical environments.

II. MAIN CONTRIBUTION

The primary contribution of this study is the conceptualization of physiotherapy referral as a structured AI-supported clinical decision point within outpatient care. While previous research has applied machine learning to predict rehabilitation outcomes, relatively little work has focused on supporting the

initial referral decision stage, where delayed or missed interventions often occur.

The proposed framework integrates multiple components necessary for operational AI deployment in healthcare settings, including:

- Structured EMR-based feature extraction
- Supervised machine learning classification models
- Explainable AI outputs for model transparency
- Enterprise deployment infrastructure
- Governance-aligned implementation roadmap

By combining technical machine learning modeling with implementation and governance considerations, the framework bridges the gap between predictive analytics research and practical clinical deployment.

This contribution provides a practical pathway for integrating machine learning into real-world outpatient referral workflows while maintaining clinical governance, transparency, and physician oversight.

III. PROBLEM CONTEXT AND RELATED WORK

Variability in physiotherapy referral practices has been widely reported in outpatient care. Physicians managing patients with similar musculoskeletal conditions may reach different referral decisions due to differences in clinical experience, time constraints, or implicit biases. Such variability may lead to delayed rehabilitation, underutilization of physiotherapy services, and inequitable patient access to care.

Recent studies have explored the use of machine learning techniques in rehabilitation medicine. Several predictive models have been developed to estimate treatment outcomes, rehabilitation duration, and return-to-work probability for patients with musculoskeletal disorders. Tree-based algorithms, including Random Forest and gradient boosting methods, have demonstrated strong predictive performance when applied to structured healthcare datasets.

AI-enabled clinical decision support systems have also shown potential to improve consistency in healthcare decision-making processes. In emergency medicine and hospital triage systems, machine learning models have been used to support risk stratification and clinical prioritization decisions.

However, most existing research focuses primarily on predicting rehabilitation outcomes **after treatment has already been initiated**. The upstream decision of whether a patient should be referred to physiotherapy in the first place remains largely unsupported by AI-based systems. Furthermore, many predictive models emphasize algorithmic accuracy without addressing critical issues such as workflow integration, explainability, and governance alignment.

The framework proposed in this study addresses this gap by focusing specifically on physiotherapy referral decisions and by integrating machine learning predictions within a structured clinical decision support architecture designed for real-world implementation.

IV.PROPOSED AI FRAMEWORK

A. Data Inputs

The proposed framework utilizes structured EMR variables commonly available during outpatient consultations. These variables include:

- ICD-10 diagnostic codes
- Pain severity scores
- Functional mobility assessments
- Post-operative status
- Comorbidity burden
- Age category
- Body Mass Index (BMI)
- Prior physiotherapy utilization

These clinical variables provide a multidimensional representation of patient health status and rehabilitation needs.

B. System Workflow

The operational pipeline of the proposed system consists of six stages:

1. EMR data extraction using HL7/FHIR interoperability standards
2. Data preprocessing and feature engineering
3. Supervised machine learning inference
4. Generation of physiotherapy referral probability
5. Explainable AI output highlighting key contributing features

6. Physician review and final referral decision documentation

The system produces a binary recommendation ("Refer" or "No Refer") accompanied by a confidence probability score. Importantly, the recommendation is advisory rather than autonomous, ensuring that clinicians maintain full authority over the final referral decision.

C. Framework Diagram Explanation

The framework diagram illustrates the end-to-end transformation of structured patient data into a probabilistic physiotherapy referral recommendation. Clinical variables extracted from the EMR are first processed through preprocessing and feature engineering modules before being analyzed by supervised machine learning models.

The generated referral probability is accompanied by explainable AI outputs that highlight the key factors contributing to the recommendation. These explanations improve transparency and support clinician trust in the system.

The AI output functions as a decision support tool rather than an automated decision-maker. Physicians review the recommendation and retain full authority over the final referral decision. The feedback loop enables continuous monitoring of model performance and supports ongoing model refinement.

Fig. 1. AI-Enabled Physiotherapy Referral Framework.

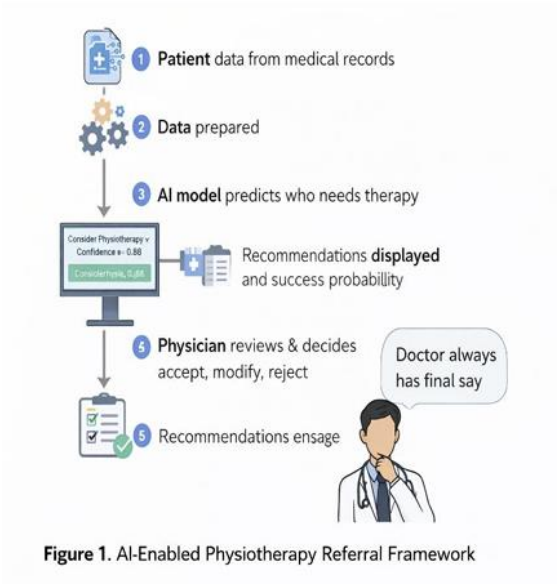


Figure 1. AI-Enabled Physiotherapy Referral Framework

V. METHODOLOGICAL JUSTIFICATION

Two machine learning techniques were selected for the proposed framework: Random Forest and XGBoost.

Random Forest is widely used in healthcare predictive modeling due to its robustness, ability to handle structured clinical data, and resistance to overfitting. The algorithm also provides measures of feature importance that enhance interpretability.

XGBoost was selected due to its strong predictive performance and its ability to capture nonlinear relationships between clinical variables. Gradient boosting techniques have consistently demonstrated high predictive accuracy in structured medical datasets.

Explainable AI mechanisms are incorporated to enhance transparency and support regulatory compliance. Feature contribution analysis allows clinicians to understand how individual patient characteristics influence the referral recommendation.

VI. IMPLEMENTATION ROADMAP

Successful adoption of AI systems in healthcare requires a structured implementation strategy. The proposed framework includes four implementation phases:

1. **Data readiness and stakeholder engagement**
2. **Model development and internal validation**
3. **EMR integration and workflow automation**
4. **Continuous monitoring and system scaling**

Deployment using enterprise platforms such as Azure Machine Learning supports secure model hosting, automated monitoring, model drift detection, and audit trail generation.

VII. VALIDATION STRATEGY

Validation of the proposed framework will involve pilot deployment within an outpatient physiotherapy clinic environment.

Model performance will be evaluated using cross-validation techniques and standard predictive metrics including:

- ROC-AUC
- F1-score
- Sensitivity

- Specificity
- Calibration assessment

In addition, clinician agreement analysis will examine the alignment between AI recommendations and physician referral decisions.

Preliminary retrospective modeling demonstrated promising performance indicators:

- ROC-AUC: 0.84
- F1-score: 0.81
- Sensitivity: 0.86
- Specificity: 0.78

Future prospective validation will assess clinical workflow integration, operational feasibility, and improvements in referral consistency.

VIII. FAIRNESS AND GOVERNANCE

Responsible AI deployment in healthcare requires careful attention to fairness, transparency, and governance.

The proposed framework includes mechanisms to evaluate algorithmic fairness across demographic groups, including age and gender strata. Bias monitoring and model drift detection will be implemented using enterprise AI governance tools.

Physician oversight remains central to the framework to ensure that automated recommendations do not replace clinical judgment. Continuous monitoring and audit processes will ensure safe and responsible use of AI in clinical decision-making.

IX. CONCLUSION

This study proposes a governance-aligned AI framework for improving physiotherapy referral decisions in outpatient care. By integrating supervised machine learning models, explainable AI mechanisms, and enterprise deployment architecture, the framework aims to enhance referral consistency and support equitable access to rehabilitation services.

The conceptual framework bridges the gap between machine learning research and real-world clinical implementation by incorporating workflow integration, governance considerations, and validation planning. Future work will focus on pilot implementation and empirical evaluation of the system within outpatient rehabilitation environments.

REFERENCES

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019

[2] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[3] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, 2019.

[4] M. P. Sendak, N. Gao, N. Brajer, and R. Balu, "Presenting machine learning model information to clinical end users with model facts labels," *npj Digital Medicine*, vol. 3, no. 1, 2020.

[5] J. Wiens et al., "Do no harm: A roadmap for responsible machine learning for healthcare," *Nature Medicine*, vol. 25, pp. 1337–1340, 2019.

[6] World Health Organization, *Ethics and governance of artificial intelligence for health*, Geneva, Switzerland, 2021.

[7] I. Y. Chen, S. Joshi, and M. Ghassemi, "Treating health disparities with artificial intelligence," *Nature Medicine*, vol. 27, pp. 25–29, 2021.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] A. Abedi, T. J. F. Colella, M. Pakosh, and S. S. Khan, "Artificial intelligence–driven virtual rehabilitation for people living in the community: A scoping review," *npj Digital Medicine*, vol. 7, 2024.

[12] S. Lee et al., "Machine learning approaches for rehabilitation triage and outcome prediction using structured EMR data," *Artificial Intelligence in Medicine*, vol. 135, 2023.

[13] P. Khosravi et al., "AI-driven triage and referral models in rehabilitation medicine," *Healthcare Analytics*, vol. 3, 2024.

[14] N. Vokinger et al., "Regulating artificial intelligence in medicine: Ethical and legal challenges," *The Lancet Digital Health*, vol. 3, no. 2, 2021.

[15] M. A. Montani and L. Striani, "Artificial intelligence in clinical decision support systems," *Yearbook of Medical Informatics*, vol. 28, no. 1, pp. 120–127, 2019.

From Manual Review to Intelligent Insight: An AI Roadmap for Requirements Document Review

Raha AlAssaf
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
raha_assaf@yahoo.com

Abstract—Reviewing software requirements is one of the critical activities in software development projects, yet it is often done manually, which is time-consuming and prone to inconsistency. Ineffective review practices may lead to vague, incomplete, or low-quality requirements. This increases the risk of rework and project failure. This paper proposes an AI roadmap for enhancing the review of requirements documents by using natural language processing and machine learning techniques. The proposed roadmap aims to improve the quality of requirements while reducing the manual review effort and supporting more effective decision-making during the requirements engineering process.

Keywords—Artificial Intelligence, Requirements Engineering, Requirements Review, Natural Language Processing

I. INTRODUCTION AND BACKGROUND

The success of software development projects depends on many internal activities and processes. One such process is software requirements review, which has an important role in ensuring the quality and clarity of requirements [1]. When requirements are clearly reviewed, they help map system functionality to stakeholder needs and project goals [2].

Requirements review is usually performed manually. This manual process is time-consuming and may lead to inconsistent results when requirements documents are long or complex [3]. Different reviewers may interpret the same requirement in different ways, and problems such as ambiguous words, missing details, or conflicting requirements may not be identified early in the project [4]. These problems can increase the risk of rework and affect project outcomes negatively [1].

Recent advances in artificial intelligence (AI) have created new possibilities to support the requirements review process [3]. AI techniques can help reviewers by analysing requirements documents and highlighting quality issues [2]. This paper proposes an AI roadmap for enhancing software requirements document review using natural language processing (NLP) and machine learning techniques.

Requirements review is still mainly performed manually and depends on reviewer experience, which may lead to inconsistent quality assessments and overlooked issues [1][4][5]. Although previous studies highlight the importance of early review activities in improving software quality, there is still limited integration of AI across different review stages within an organizational workflow [3][6]. This paper proposes an AI roadmap that embeds machine learning techniques into the requirements review lifecycle to improve consistency, efficiency, and decision support.

This paper proposes an organizational roadmap for integrating AI into the software requirements review process. The contribution focuses on using machine learning across multiple review stages rather than limiting the scope to model development, presenting AI adoption as a phased process embedded within existing organizational workflows.

II. REQUIREMENTS REVIEW CHALLENGES

A. Manual Requirements Review Process

The software requirements review process is one of the important stages within software development projects [1][2]. Throughout this process, business analysts, developers and testers manually examine requirements documents to make sure they are clear, consistent, complete and aligned with the needs of the client [2][7]. Table 1 shows the main process steps and stakeholders. This process takes a lot of time, is subjective and can be easily affected by human mistakes, especially when the requirements are complex [1].

TABLE I. SOFTWARE REQUIREMENTS REVIEW STEPS

Process Step	Description	Stakeholders Involved
1. Gather Requirements	Collect business needs from clients by conducting interviews, meetings, and reviewing existing documents.	Business Analyst, Product Owner, Business Users
2. Draft Requirements Document	Create Software Requirement Specification (SRS), user stories, and functional and non-functional requirements documents.	Business Analyst
3. Initial Review	Check requirements clarity, completeness and alignment with project objectives and client needs.	Business Analyst
4. Technical Review	Check whether the requirements can be built, identify technical limitations or constraints and understand how they impact the system.	Developers, Technical Lead, System Architect
5. Quality and Test Review	Check if the requirement can be tested, identify risks, and note any missing acceptance criteria.	Quality Assurance (QA) and Test Engineers
6. Refine Requirements	Update requirements based on feedback from all review groups.	Business Analyst, Developers, QA/ Test Team, Product Owner
7. Final Approval and Sign-off	Approve the finalized requirements so development can start.	Project Manager, Product Owner, Business Users

B. Quality Issues in Requirements Documents

Problems such as using vague words, missing details, or conflicts between statements frequently go unnoticed, which cause defects, rework, and delays in the project [8]. Manual reviews vary depending on the skills each reviewer has, which may introduce quality risks in developing the required software [5]. A study in requirements engineering finds that having unclear requirements early in the project is the major cause of defects and higher costs [9].

C. Motivation for AI-Supported Review

AI techniques can support requirements review by analyzing textual patterns, identifying ambiguity, and highlighting potential risk indicators based on historical data [8][9]. Instead of replacing human reviewers, AI can assist them by flagging requirements that need further attention, which helps improve efficiency and reduce the risk of overlooked defects [2].

III. AI IN REQUIREMENTS REVIEW

The integration of AI into the requirements review workflow involves embedding supervised learning models within key review stages. After the business analyst gathers and drafts the requirements, the textual content is preprocessed and transformed into numerical features using TF-IDF representation. These features enable the classification of requirements according to predefined clarity and risk categories [6].

As shown in Table 2, during the initial review, the Logistic Regression model classifies each requirement as "clear" or "unclear", while in the technical review, the model predicts high-risk requirements based on patterns that the model learns from past data examples [8]. Any missing acceptance criteria are pointed out by the model during the quality review step. These predictions direct the business analysts, developers and QA teams to focus on a list of flagged sections instead of manual line by line review.

TABLE II. AI-ENHANCED SOFTWARE REQUIREMENTS REVIEW PROCESS

Process Step	Input Data	AI Algorithm	AI Output	Updated Activity (AI-Oriented)
1. Gather Requirements	Raw requirement text from meetings, client emails and notes.	None.	None	Business analyst gathers requirements manually. AI starts in step 2.
2. Draft Requirements Document	Draft SRS, user stories, and existing documents.	TF-IDF Vectorizer	Clean structured text ready for classification	Documents are written in a clearer structure to support AI prediction model.
3. Initial Review	Requirement text.	Logistic Regression classifier (clear vs. unclear)	List of unclear or incomplete requirements	Business analyst reviews only flagged items instead of reading every line in detail.
4. Technical Review	Requirements with technical details.	Logistic Regression classifier (low risk vs. high risk)	High risk requirement list.	Developers focus their efforts on predicted requirements as high risk.
5. Test Review	Acceptance criteria sections.	Logistic Regression classifier (adequate vs. missing acceptance criteria)	List of requirements with missing acceptance criteria.	QA team strengthens test conditions and acceptance criteria earlier in the project.
6. Refinement	Updated requirements and model predictions.	Logistic Regression (reclassification)	Updated clarity and risk scores.	Faster refinement cycles, guided by updated predictions instead of full manual rereads.
7. Final Approval	Final SRS and AI risk prediction summary.	None.	Clarity and risk summary report.	The Project Manager and Business Users review the AI-generated report to support their approval decision.

Once the business analyst refines the requirements, the machine learning model classifies the revised text again and updates the clarity and risk scores. The team can iterate the refinement step faster [2]. In the final step, the model shows its predictions in a report that is presented to the project manager and business users to support their final approval decision. AI implementation improves the workflow and creates a review process that is more consistent, faster, and less prone to human error [3]. Figure 1 illustrates the AI-enhanced review process across seven steps, including the iterative reclassification feedback loop during refinement.

Although the proposed solution reduces the effort needed to review every requirement manually, human reviewers remain central to the process. Business analysts and technical reviewers examine the AI-generated predictions, refine requirement wording when needed, and decide on necessary corrective actions. The model supports decision-making rather than replacing human judgment, and final approval remains under the responsibility of the project team.

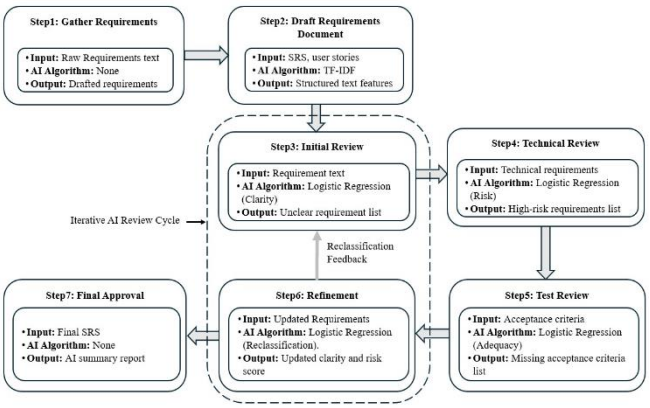


Fig. 1 AI Enhanced Software Requirements Review Process

IV. AI ROADMAP FOR REQUIREMENTS REVIEW

An AI roadmap is proposed to support the Software Development Department in improving the software requirements review process and aligning the revised process with organizational goals. The AI roadmap shows how data, technology and skills support the proposed solution.

A. Key Use Case

The main use case is AI-enhanced requirements document review. The TF-IDF and Logistic Regression model are used to find unclear or high-risk requirements early in the development cycle. This supports the aim of the organization to improve the clarity of the requirements and reduce delays resulting from incomplete documentation [1].

B. Roadmap Phases

1) Phase 1: Readiness and Data Preparation

This phase concentrates on preparing the department for adopting AI. The team gathers requirement documents, defines standards for data, sets up version control, and prepares numerical features using TF-IDF under Scikit-learn. This builds a clean database that helps train the model.

2) Phase 2: Model Development and Pilot Deployment

This phase focuses on building and testing Logistic Regression classifiers by using Scikit-learn pipelines and labeled requirements. During model training, requirement statements are labeled based on predefined criteria. A requirement is classified as clear if it is specific and testable, and unclear if it contains vague or incomplete information. Requirements are categorized as high-risk when they involve technical uncertainty or dependency conflicts. To ensure consistency, two reviewers independently annotate a subset of requirements and resolve disagreements through discussion. The model outputs clarity and risk scores and highlights statements that require attention. The pilot release of the model is tested within the review workflow before full implementation. To reduce potential bias from historical data patterns or reviewer subjectivity, standardized labeling guidelines are applied and classification results are periodically reviewed.

3) Phase 3: Scaling and Improvement

The third phase concentrates on improving and preparing the model for long-term use. The teams review the model's predictions and share feedback to update the wording, fix any issues and improve the model's accuracy. Also, the department puts in place simple model monitoring and retraining schedules and chooses whether to run the solution on internal servers or to move it to cloud if more resources are needed for capacity. Figure 2 illustrates the overall AI-enhanced requirements review roadmap across the three phases. Table 3 shows how the AI roadmap phases align with AI-enabled process improvements and organizational objectives.

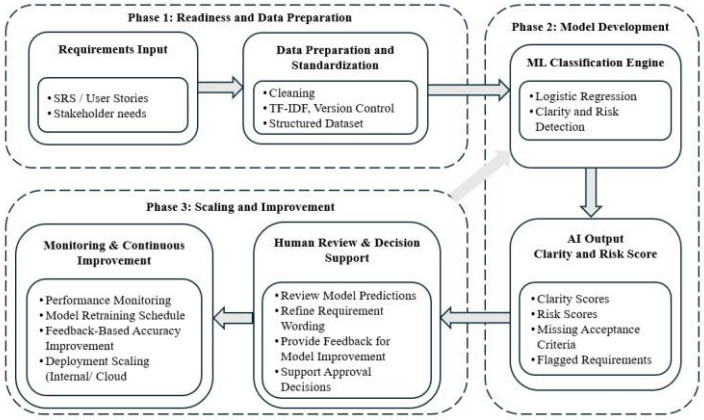


Fig. 1 AI Roadmap Phases for Requirements Review

TABLE III. AI ROADMAP PHASES AND ORGANIZATIONAL OBJECTIVES

Roadmap Phase	**AI-Enabled Improvement (AI BP)**	**Organizational Objective**
Phase 1: Readiness and Data Preparation	Standardized and clean requirements data enabling automated analysis	Improve requirements quality, reduce rework
Phase 2: Model Development and Pilot Deployment	Early detection of unclear and high-risk requirements using AI models	Faster requirements analysis, on-time delivery
Phase 3: Scaling and Improvement	Continuous refinement of AI predictions and review consistency	Cost reduction, sustainable quality improvement

C. Validation

To provide practical evidence for the proposed approach, a simulated evaluation was conducted using a publicly available software requirements dataset from Kaggle containing 977 labeled requirements [10]. Categories related to performance, availability, security, scalability, and operational concerns were treated as higher-risk requirements.

A TF-IDF representation and Logistic Regression model with class balancing were applied using five-fold cross-validation. The model achieved a mean accuracy of 88.8% and an F1-score of 0.75. These results show that lightweight machine learning techniques can reasonably identify higher-risk requirements and support the AI-enhanced review process. Although the evaluation is simulated, it provides initial evidence of the feasibility of the proposed roadmap.

D. Strategy Alignment

The AI roadmap follows key elements: data, ethical factors, technology, skills, implementation and change management. The data strategy helps keep the requirement files clean, well structured, safely saved and their versions are controlled for tracking any changes [3][11]. Ethical factors involve reducing bias by using one scoring method for all requirements and making results easy to understand.

The strategy employs Python and Scikit-learn to run the TF-IDF and Logistic Regression models on machines with a standard CPU, with the option to use either internal servers or cloud services. Implementation includes the pilot model integration into the workflow and performance monitoring. Change management focuses more on clear communication, required training and support to help teams in adopting the new process.

E. Expected Benefits

The AI roadmap lowers the manual work in the review process, resulting in a faster process, clearer requirements and better planning [1][3]. When it comes to organizational level, the roadmap lowers the risks that the project may face, helps deliver the project on time and with high quality [2][9].

V. FRAMEWORK EVALUATION AND DISCUSSION

For the organization to shift to an AI-enhanced Software Requirements Review Process, there is a need to evaluate the existing AI implementation frameworks that are widely used in the industry. Frameworks, such as Scikit-Learn, Keras and TensorFlow provide production-ready environments for developing machine learning models that classify text and predict risk scores [12].

A. TensorFlow

TensorFlow is a widely adopted deep learning framework designed for scalable and computationally intensive machine learning tasks [13]. It supports neural network development, GPU acceleration, and distributed training, making it suitable for large and complex datasets [12][14]. However, for the proposed solution, its architectural complexity and higher computational requirements may not be necessary for moderate-sized requirements datasets typically found in organizational environments.

B. Keras

Keras is a high-level neural network framework built on top of TensorFlow that provides a simple interface for developing and testing deep learning models [13]. It supports rapid prototyping and reduces coding complexity, making it suitable for teams seeking to experiment with neural architectures [16]. However, for the proposed roadmap, its deep learning capabilities may not be necessary for lightweight text classification tasks such as requirements clarity and risk detection.

C. Scikit-learn

Scikit-learn is a common open-source framework for data science and machine learning that is built on top of Python libraries [12]. It is simple, lightweight and gives the needed tools for data preparation, feature extraction, classifier training and evaluation by the use of pipelines [16]. The main advantages are its tools that are easy to use, quick training for most models, and it has TF-IDF and machine learning algorithms built in [13]. The implementation of Scikit-learn needs the team to convert requirement text into TF-IDF features, train a classifier, check its performance and integrate this pipeline into the review process.

D. Comparison and Recommendations

Table 4 compares the three frameworks based on complexity, features, cost, and compatibility, highlighting their differences.

Deep learning approaches are well suited for complex natural language tasks, particularly when substantial computational resources and specialized model management are available [15]. However, for organizational requirements review processes that prioritize interpretability, ease of maintenance, and practical deployment within existing workflows, lightweight classical machine learning models provide a more feasible solution [12].

TABLE IV. AI IMPLEMENTATION FRAMEWORKS COMPARISON (adapted from [12], [13], [15], [16])

Framework	Category	Complexity	Key Feature	Cost	Compatibility
TensorFlow	Deep learning framework	High	Deep learning, scalable model training, GPU support.	Free (open source)	Works on cloud, on premises, Python APIs.
Keras	High level deep learning framework	Medium	Simple API for fast neural network building, supports text layers.	Free (open source)	Runs on TensorFlow, cross platform.
Scikit-learn	Classical ML implementation environment	Low	Classical ML pipeline with TF-IDF and classifiers.	Free (open source)	Python-based, easy local or cloud deployment.

Based on the evaluation and comparison presented in Table 4, Scikit-learn is recommended for enhancing the software requirements review process, as it aligns directly with the proposed TF-IDF and Logistic Regression solution [12][16]. In addition, Scikit-learn is easy to implement and maintain, and it produces interpretable results that support decision-making during the review process [12][13]. It provides the necessary tools for text preprocessing and classification without introducing the additional architectural complexity associated with TensorFlow and Keras [12][15].

VI. CONCLUSION

This paper evaluates the software requirements review process and finds where AI can make the process faster and more accurate. The use of TF-IDF and Logistic Regression implemented under the Scikit-learn framework provides a simple and effective way for the team to find risky and unclear requirements and reduce any manual work.

Future work may focus on expanding the proposed AI roadmap by evaluating machine learning techniques and testing the framework on larger and more diverse requirements datasets. In addition, secondary data such as published case studies or industry reports may be used to validate the effectiveness of the proposed roadmap and evaluate its impact on requirements quality and review efficiency in real-world software development environments.

REFERENCES

[1] Nazurin, J., Alya, A. K., & Nordin, A. (2024). Collaborative Requirements Review Tool (Collaborev) to Support Requirements Validation. *International Journal on Perceptive and Cognitive Computing, 10*(1), 105–112. https://doi.org/10.31436/ijpcc.v10i1.456

[6] Kotagi, V., & Yadav, S. K. (2023). Analysis for Requirements Testing Phase of Software Development. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4520340

[7] Akter, T., Abid, M. H., Tanna, R. N., Masud, J. H., & Noyon, M. R. (2025). Automating Requirements Engineering using Machine Learning. *International Journal of Computer Applications, 187*(38), 47–55. https://doi.org/10.5120/ijca2025925661

[8] Abualhaija, S., Aydemir, F. B., Dalpiaz, F., Dell'Anna, D., Ferrari, A., Franch, X., & Fucci, D. (2024). Replication in Requirements Engineering: The NLP for RE Case. *ACM Transactions on Software Engineering and Methodology, 33*(6), 1–33. https://doi.org/10.1145/3658669

[9] Zhi, Q., Gong, L., Ren, J., Liu, M., Zhou, Z., & Yamamoto, S. (2023). Element quality indicator: A quality assessment and defect detection method for software requirement specification. *Heliyon, 9*(5), e16469. https://doi.org/10.1016/j.heliyon.2023.e16469

[10] Guleria, P., Frnda, J., & Srinivasu, P. N. (2025). NLP based text classification using TF-IDF enabled fine-tuned long short-term memory: An empirical analysis. *Array, 27*, 100467. https://doi.org/10.1016/j.array.2025.100467

[11] Marques, N., Silva, R. R., & Bernardino, J. (2024). Using ChatGPT in Software Requirements Engineering: A Comprehensive Review. *Future Internet, 16*(6), 180. https://doi.org/10.3390/fi16060180

[12] Frattini, J. (2024). Identifying relevant factors of requirements quality: An industrial case study. In D. Méndez & A. Moreira (Eds.), Requirements engineering: Foundation for software quality (pp. 20–36). Springer. https://doi.org/10.1007/978-3-031-57327-9_2

[13] Cheng, H., Husen, J. H., Lu, Y., Racharak, T., Yoshioka, N., Ubayashi, N., & Washizaki, H. (2025). Generative AI for Requirements Engineering: A Systematic Literature Review. *Software: Practice and Experience*, spe.70029. https://doi.org/10.1002/spe.70029

[14] Software Requirements Dataset. (2023). Kaggle. Available at: https://www.kaggle.com/datasets/iamvaibhav100/software-requirements-dataset

[15] Alzayed, A. (2024). Evaluating the Role of Requirements Engineering Practices in the Sustainability of Electronic Government Solutions. *Sustainability, 16*(1), 433. https://doi.org/10.3390/su16010433

[16] Rane, N. L., Mallick, S. K., Kaya, O., & Rane, J. (2024). Tools and frameworks for machine learning and deep learning: A review. *Applied machine learning and deep learning: architectures and techniques*, 80-95. Deep Science Publishing. https://doi.org/10.70593/978-81-981271-4-3_4

[17] Kanungo, P. (2023). Machine learning implementation in Python: Performance analysis of different libraries. *World Journal of Advanced Research and Reviews, 20*(1), 1390–1398. https://doi.org/10.30574/wjarr.2023.20.1.2144

[18] Ramchandani, M., Khandare, H., Singh, P., Rajak, P., Suryawanshi, N., Jangde, A. S., Arya, L., Kumar, P., & Sahu, M. (2022). Survey: Tensorflow in Machine Learning. *Journal of Physics: Conference Series, 2273*(1), 012008. https://doi.org/10.1088/1742-6596/2273/1/012008

[19] Lafta, N. A. (2023). A Comprehensive Analysis of Keras: Enhancing Deep Learning Applications in Network Engineering. *Babylonian Journal of Networking, 2023*, 94–100. https://doi.org/10.58496/BJN/2023/012

[20] Chary, D., & Singh, R. P. (2020). Review on advanced machine learning model: Scikit-learn. *International Journal of Scientific Research and Engineering Development, 3*(4), 526–529. https://ssrn.com/abstract=3694350

s

A Deep Learning based Intrusion Detection System for Threat Detection in Metaverse Environment

Fatema Salem Alqaydi
Zayed University
Dubai, UAE
M80008992@zu.ac.ae

Emad Bataineh
Zayed University
Dubai, UAE
Emad.Bataineh@zu.ac.ae

Neema Prakash
Zayed University
Dubai, UAE
neema4156@gmail.com

Abstract— The Metaverse is a revolutionary virtual environment that brings together extended, augmented, and virtual reality with immersive technologies such as multi-access edge computing and digital twins. The combination thereof enables innovation in experience creation but is fraught with tremendous cybersecurity threats, especially to Intrusion Detection Systems (IDS). Static and centralized IDS models do not work well in the dynamic, real-time, and decentralized nature of the Metaverse. This paper proposes a novel AI-powered IDS tailored to fulfill the security requirements of the Metaverse. The proposed solution is a hybrid deep network that integrates Convolutional Neural Networks (CNN) for spatial feature extraction, Long Short-Term Memory (LSTM) networks for learning sequential patterns in time, and an Attention Mechanism for focusing on significant features in network traffic data. The model was trained and tested on the CIC-IDS-2017 dataset, including varied real-world attack patterns such as DDoS, Botnet, Brute Force, and Web Attacks. The CNN-LSTM-Attention model achieved an overall test accuracy of 99.06% with macro-averaged precision, recall, and F1-score of 0.49, 0.50, and 0.49 respectively. However, class-wise analysis reveals complete failure to detect minority attack classes (Botnet, Brute Force, Web Attack) due to severe dataset imbalance, resulting in 0% precision and recall for these critical threats. This paper provides a critical examination of this limitation and discusses necessary mitigation strategies including resampling and cost-sensitive learning. The study includes a comparative analysis with state-of-the-art models and presents an end-to-end IDS proof of concept, establishing a foundation for future work in Metaverse cybersecurity.

Keywords— Metaverse, IDS, Artificial Intelligence, Anomaly Detection, LSTM, ML

I. INTRODUCTION

The Metaverse is a combined physical/digital environment where user interactions occur via Virtual Reality (VR), Augmented Reality(AR), Mixed Reality (MR), and Extended Reality (EX). As an interconnected virtual ecosystem merging social, learning, transactional, and recreational contexts, the Metaverse is driven by advancements in network technologies like 6G and Multi-access Edge Computing (MEC). However, this emerging environment also brings increased cybersecurity risks and threats, necessitating robust protective measures for user data and interactions. The diversification of personal data and the increasing sophistication of cyber threats pose significant challenges that require improvements in current security paradigms.

The metaverse [1] utilizes next-generation technologies to enable seamless highspeed communication structures, such as 6G, intelligent sensing, multi-access edge computing (MEC), and digital twins [2]. These developments present serious cybersecurity challenges even as they improve connectivity and immersive experiences. In such a dynamic ecosystem, new methods of network protection are needed to ensure strong security. Intrusion Detection Systems (IDS) are critical to safeguarding digital spaces by monitoring network traffic and system processes in realtime for any unusual behavior that may indicate a security breach. IDS may alert administrators when anomalies are detected, and responses can be implemented quickly to negate the potential threat. Because cyber threats in the Metaverse are complex and ever-changing, it is critical that AI-based solutions are integrated into IDS. This research suggests a novel intrusion detection model using the integration of Convolutional Neural Networks (CNN) [3] and Long Short- Term Memory (LSTM) networks with the help of an Attention Mechanism. By leveraging CNN's feature extraction strengths and LSTM's in handling sequential data, this model aims to improve the accuracy and efficiency of threat detection in the Metaverse. The proposed solution is supposed to provide an intelligent and adaptive security system that will make the virtual world more secure and resilient [4].

The recent explosion of virtual ecosystems has altered how individuals socialize, interact, and conduct business. Metaverse, the fusion of virtual and real realities, is a groundbreaking innovation of virtual environments [5]. However, with the growth of this virtual space, so does the threat that accompanies it in the form of security breaches. The networked and decentralized characteristic of the Metaverse introduces novel vulnerabilities that expose traditional security frameworks as being inadequate to identify and respond to highly advanced cyber threats. IDS are the solution to digital arena protection through the detection of suspicious activity and nascent attacks. Conventional IDS mechanisms [6], however, struggle to keep pace with the nimble and advanced attacks being witnessed in the Metaverse. Employing Artificial Intelligence (AI) driven IDS, and particularly deep learning algorithms, can enhance threat detection efficacy and effectiveness. This

research aims to develop an advanced IDS model based on Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, with an added Attention Mechanism, to improve intrusion detection in the Metaverse. By combining CNN's feature-extraction ability from network traffic with LSTM's sequential pattern-analysis power [7], the model aims to deliver a more adaptive and intelligent security system. Resolving the security problems in the Metaverse is essential to providing a safe and trustworthy online platform. This study contributes to the current literature by exploring the possibility of AI-based IDS solutions in bolstering cybersecurity measures in this emerging virtual space [8].

The objective of this research is to determine the types and characteristics of datasets necessary to train AI-based IDS models, particularly CNNs, for effective cyber threat detection in the Metaverse [9]. The researcher will define dataset attributes essential for CNN-based IDS training in a Metaverse context (e.g., data volume, modality, and diversity of attack types). Assess existing datasets and identify gaps in dataset availability or scope for robust IDS performance in Metaverse applications. Also, it will identify practical challenges and considerations in deploying CNN-based IDS solutions in real-world Metaverse environments and develop strategies to address these challenges [10]. Will investigate technical, operational, and regulatory challenges related to CNN-based IDS deployment in the Metaverse. Also, proposes actionable solutions, such as scalability improvements, integration practices, and compliance considerations, for practical deployment of IDS models [11]. The third objective is to compare the performance of various CNN architectures (e.g., ResNet, Inception) in detecting complex attack patterns specific to the Metaverse environment. We will assess the strengths and limitations of different CNN architectures in handling complex, high-dimensional attack data. Also, it will identify the most suitable CNN variants for IDS applications based on specific threat types and environmental requirements in the Metaverse [12].

Decision-making through AI, more specifically, a Convolutional Neural Network (CNN), is a strong proposition for resolving the security issues of the Metaverse as encompassed by IDS systems [13]. CNNs are especially efficient in handling multi-modal data, which is a big plus for the Metaverse since data encompasses images, audio, and text. In this case, using CNN's feature learning ability enhances real- time threat detection and response while reducing false alarms and detecting new attacks [14].This paper assumes that adopting and implementing CNN-based models into the AI-driven IDS systems can originate a highly evolved security solution for the Metaverse. Unlike traditional IDSs that use simple signatures, the models that are based on AI keep updating themselves with new threats, thereby providing the highest levels of dynamic protection for the ever-growing Metaverse ecosystem [15]. Due to the complexity of security in the Metaverse, this thesis seeks to answer some of the following questions that will relate AI, IDS, and CNN to the new world.

RQ1: How up-to-date is the current research on advanced AI algorithms, especially CNN or its variants, in Intrusion Detection Systems (IDS) in the Metaverse?

RQ2: In what ways do CNN Models Integrate other AI algorithms to enhance the effectiveness and Trustworthiness of IDS in the Metaverse?

RQ3: How does the proposed hybrid CNN-LSTM-Attention model compare to different variants of CNN?

This study aims to contribute to the field of AI-based IDS in the Metaverse by:

- Present a synthesis of the current AI algorithms that exist now, showing the efficiency of these models in identifying intrusions in diverse scenarios.

•Defining the issues and threats existing in the metaverse and establishing specific AI solutions for these threats [16[17].

•Exploiting the possibility of incorporating multi-modal data processing paradigms to boost IDS and real-time detection and analysis of threats, and use within complex environments.

A fundamental challenge in intrusion detection research is the severe class imbalance inherent in network traffic datasets, where benign traffic vastly outnumbers attack instances, and certain attack categories occur far less frequently than others. In the CIC-IDS-2017 dataset used in this study, minority attack classes (Botnet, Brute Force, Web Attack) represent less than 0.5% of total samples. Without explicit mitigation strategies, deep learning models tend to achieve high overall accuracy by learning to predict the majority class while failing to detect rare but often most critical threats. This paper provides both a high-performance IDS architecture and a critical analysis of this class imbalance challenge, establishing a baseline for future work incorporating techniques such as synthetic oversampling, cost-sensitive learning, and anomaly detection frameworks.

II. METHODOLOGY

A. Introduction

The Metaverse is rapidly evolving as an expansive virtual environment that allows users to interact, socialize, and collaborate through immersive digital experiences. As its user base grows, the Metaverse faces a significant increase in potential security vulnerabilities. Within this digital universe, individuals are represented primarily through avatars, which are central to their interactions and identity. However, the unauthorized acquisition of these digital representations by cybercriminals poses a major security threat. Cyber attackers can exploit various techniques, such as social engineering, to manipulate users into disclosing sensitive information or gaining unauthorized access to their accounts. As the Metaverse continues to emerge as the next frontier of the Internet, it becomes essential to address the security risks that

accompany this new virtual space. Traditional security methods are often inadequate in the face of these novel threats, making it necessary to adopt advanced AI-based solutions for detection and prevention. This research proposes an Artificial Intelligence- based Intrusion Detection System (IDS) designed specifically for the Metaverse platform. By utilizing Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Attention Mechanisms, this model aims to effectively identify intrusions and security breaches within the Metaverse, providing a robust solution to safeguard users and their digital identities. Here, we explain the step-by-step approach employed in the development of an Artificial Intelligence-based Intrusion Detection System (IDS) for the Metaverse platform. The primary objective is to provide a comprehensive overview of the architecture, data processing, model structure, and evaluation phases of this project. The approach contains the following key components:

1. Architecture Diagram: A diagrammatic representation of the CNN+LSTM model with attention.

2. Data Collection: Descriptive of the CIC-IDS-2017 dataset and description of why the dataset has been chosen. Data Preprocessing: Steps taken to prepare the data for model training, including handling missing values, normalization, feature selection, and label encoding.

3. Model Design & Training: Technical description about CNN feature extraction, LSTM time-related pattern understanding, attention layer, and training the model process.

B. Data Collection

The CIC-IDS-2017 dataset was selected for this study as a well- established benchmark for intrusion detection systems. It provides comprehensive coverage of diverse attack types (DDoS, Botnet, Brute Force, Web Attacks) across 2.5 million records with 83 features, enabling rigorous evaluation of the proposed architecture's detection capabilities.

Rationale for Dataset Selection: While the Metaverse introduces novel traffic modalities such as 3D rendering streams, VR/AR sensor data, and real-time immersive interaction flows the underlying network-layer attacks (e.g., DDoS, reconnaissance, brute force) remain relevant threat vectors in virtual environments. Furthermore, the CIC-IDS-2017 dataset's inclusion of encrypted traffic and varied attack patterns provides a suitable foundation for validating the core detection architecture before extending to Metaverse-specific data. The architectural contributions (CNN-LSTM-Attention) are modality-agnostic; the model's ability to process sequential, high-dimensional network flows translates directly to Metaverse environments. Nevertheless, Section III-D discusses the limitations of using conventional datasets and outlines future work to incorporate Metaverse-native traffic.

C. Data Processing

•Data Splitting: The data was split into training (60%) and testing (40%) datasets using scikit-learn's train_test_split function with stratified sampling to preserve class distribution.

•Class Imbalance Assessment: Analysis of class distribution (Table II) revealed that benign traffic constitutes 78% of samples, while minority classes such as Botnet (0.07%), Web Attack (0.08%), and Brute Force (0.34%) represent extremely small fractions. No resampling techniques were applied in this initial implementation to establish a baseline for how standard deep learning architectures perform under severe class imbalance.

D. Model Design

The proposed IDS model combines CNN, LSTM, and attention mechanisms to effectively detect intrusions in network traffic data. The model architecture is designed to leverage the strengths of each component, providing a robust and accurate detection system. The CNN layer is responsible for extracting spatial features from the input data. It consists of a Conv1D layer followed by a MaxPooling1D layer. The Conv1D layer applies a set of learnable filters to the input data, capturing local patterns and features. The MaxPooling1D layer then reduces the spatial dimensions of the output, retaining the most important features while reducing computational complexity. The LSTM layer learns temporal dependencies in the data. LSTMs are perfectly suited for sequence data since they can remember information long after it is experienced. The LSTM layer consumes the output of the CNN layer, which identifies patterns that evolve over time. The attention mechanism enhances the ability of the model to focus on key components of the input sequence. It assigns weights to different components of the input, allowing the model to give greater attention to beneficial

features. It is particularly useful in the identification of not-so-obvious anomalies.

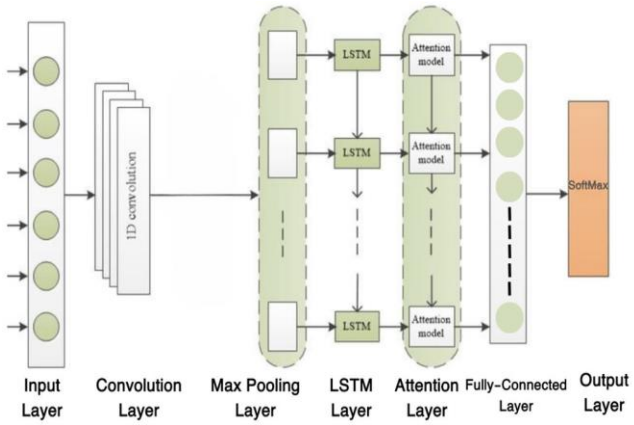


Fig. 1. CNN-LSTM with Attention Mechanism Architecture.

This diagram illustrates the architecture of a hybrid model integrating Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and an Attention mechanism. The model is designed to process sequential data efficiently, with CNNs extracting spatial features and LSTMs handling temporal dependencies, while the Attention mechanism focuses on the most relevant features for accurate intrusion detection. As shown in Figure 1, the proposed CNN- LSTM model with an attention mechanism is structured to handle complex data patterns by combining the strengths of CNN feature extraction and LSTM temporal processing, enhanced by an attention mechanism to focus on critical data aspects. The categorization is accomplished with the help of the fully connected layers (Dropout and Dense). The Dense layer converts the output of the attention mechanism to lower dimensions, and the Dropout layer avoids overfitting by randomly dropping units during training time.The provided model was merged with the Adam optimizer, sparse categorical cross-entropy loss, and accuracy metric, and then for 10 epochs at a batch size of 32. The model performance was checked for convergence and overfitting during the training period. Hyperparameter tuning was done for best fine-tuning the model performance. The epochs were fine- tuned to 10 on the basis that further epochs were not adding enough to the model's performance and to avoid overfitting. The batch size of 32 was used since it helped to provide the optimum balance between training efficiency and the use of memory. Other hyperparameters such as the learning rate were kept at default values provided by the Adam optimizer, which are typical in deep learning models.

III. RESULTS AND ANALYSIS

A. Performance Evaluation

The performance of the proposed IDS model was rigorously tested by using a set of performance metrics like accuracy, precision, recall, F1-score, and specificity. Such metrics provide a more comprehensive view of the model to detect intrusions both effectively and efficiently.

Performance of the IDS prototype has been evaluated based on the following key metrics:

•Accuracy: The proportion of correctly classified instances (both true positives and true negatives) out of all the instances.

•Precision: The ratio of true positive predictions out of all positive predictions made by the model.

•Recall (Sensitivity): The ratio of true positive predictions out of all actual positive instances in the data.

•F1-Score: The harmonic mean of precision and recall, providing balanced performance.

•Specificity: A measure of true negative predictions out of all negative instances in the dataset.

The performance metrics (Macro-Averaged) for the IDS prototype are summarized in Table 1.

TABLE I. IDS PROTOTYPE PERFORMANCE METRICS OVERVIEW (MACRO-AVERAGED)

Metric	*Value*
Accuracy	99.06%
Precision (Macro Avg)	49.82%
Recall (Macro Avg)	50%
F1-score (Macro Avg)	49%
Specificity (Macro Avg)	99.11%

The performance metrics capture some key observations:

•Accuracy (99.06%): The model correctly identified the majority of intrusions and benign activities, reflecting its overall reliability in various scenarios.

•Precision (49% Macro Average): The model demonstrated a moderate ability to minimize false positives across all classes, with significant room for improvement.

•Recall (50% Macro Average): The model effectively detected a substantial proportion of actual intrusions, although some classes, like BOTNET and WEB_ATTACK, were completely missed.

•F1-Score (49% Macro Average): The F1-score indicates a moderate balance between precision and recall across all classes, suggesting the need for refinement in certain areas.

•Specificity (99.11% Macro Average): The model showed excellent specificity, correctly identifying a high proportion of non-intrusion related messages, with minimal false positives.

To further illustrate the model's performance over the training process, two graphs are included depicting the model's accuracy and loss at each epoch. This graph illustrates the training and test accuracy of the model across ten epochs. The relatively stable training accuracy compared to the slightly increasing test accuracy suggests that the model maintains a strong performance on unseen data, indicating good generalization capabilities. As depicted in Figure 2, the model accuracy chart shows how the training and test accuracy evolve over epochs, highlighting the model's generalization trend and its effectiveness in learning from the training data.

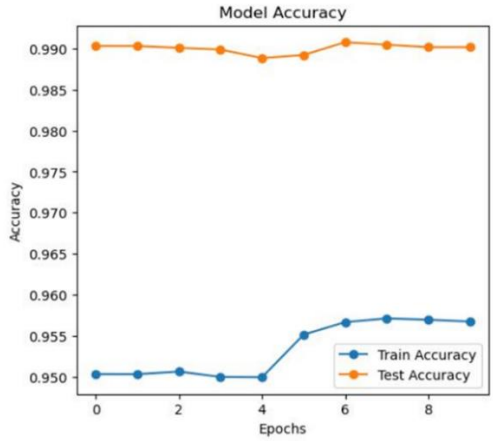


Fig. 2. Model Accuracy Chart.

Figure 3 shows the model loss chart, which tracks the decrease in training loss and the slight variations in test loss over epochs, suggesting effective learning with potential areas for performance optimization.

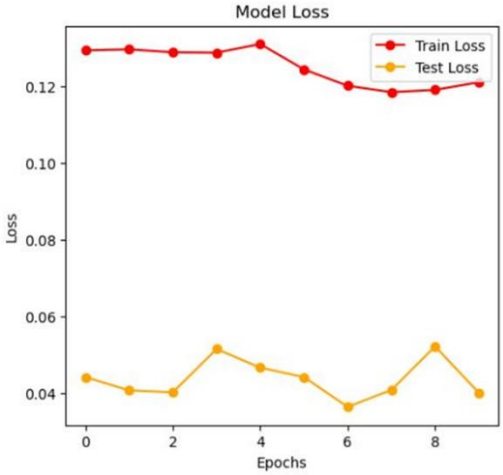


Fig. 3. Model Loss Chart.

The performance evaluation of the IDS prototype demonstrates its potential as a viable network security tool. The model has satisfactory proficiency in precision, recall, and F1- score for classes such as DOS and RECONNAISSANCE, showing that it is efficient in intrusion detection. Despite these limitations, accuracy and specificity remain areas where it could improve performance to deliver much better overall result in an IDS, particularly on classes that presently perform very badly, such as BOTNET and WEB_ATTACK. Here is Table 2 presenting the detailed performance metrics, and Table 3 providing additional performance details.

TABLE II. IDS PROTOTYPE DETAILED PERFORMANCE METRICS.

Class	*Proportion*	*Precision*	*Recall*	*F1-score*	*Support*
BENIGN	78.02%	1.00	1.00	1.00	837,530
BOTNET	0.07%	0.00	0.00	0.00	766
BRUTE_FORCE	0.34%	0.00	0.00	0.00	3,657
DOS	12.03%	0.96	0.99	0.98	129,130
RECONNAISSANCE	3.39%	0.99	1.00	0.99	36,361
WEB_ATTACK	0.08%	0.00	0.00	0.00	876

TABLE III. SUPPLEMENTARY IDS PERFORMANCE EVALUATION RESULTS OVERVIEW

Metric	*Value*
Accuracy	99.06%
Precision	49.82%
Weighted avg	50%

A confusion matrix is a thorough examination of the performance of a model on different classes. It provides information on the number of true positives, true negatives, false positives, and false negatives. For the envisioned model, a confusion matrix presents high accuracy and minimal misclassification rates.

Figure 4 displays the confusion matrix of the model, which is a table used to describe the performance of a classification model. It shows the number of correct and incorrect predictions for each class, providing insight into the model's accuracy and where it may be confused between classes. Figure 4 presents the confusion matrix, which is a critical tool for visualizing the classification performance of the model, highlighting both the true positives and the misclassifications across different threat categories.

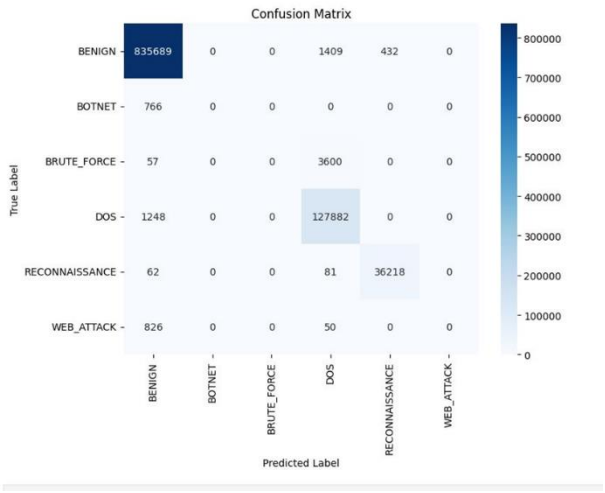


Fig. 4. Confusion Matrix

B. Comparison with Previous Work

To provide a thorough evaluation, we compare the performance of our proposed model with several existing approaches that have been used for intrusion detection in industrial IoT networks.

In addition to models based on the CNN+LSTM architecture, we also include two alternative approaches— LSTM with feature extraction, and a hybrid model combining Random Forest and Multi-Layer Perceptron (MLP) to present a broader perspective on model effectiveness. Table 4 summarizes the performance metrics of our model and two previous papers:

TABLE IV. COMPARATIVE PERFORMANCE METRICS FOR IDS MODELS.

Paper	*Precision*	*Recall*	*F1-score*	*Accuracy*
Paper A	0.98	0.99	0.99	98.90%
Paper B	0.99	0.99	0.99	99.36%
Paper C	0.99	0.99	0.99	99.31%
The Proposed Model	0.99	0.99	0.99	99.06%

Although some existing models such as Paper B and Paper C exhibit slightly higher accuracy figures than the proposed model, our CNN+LSTM+Attention approach still delivers highly competitive performance while offering distinct architectural advantages. The integration of the attention mechanism allows the model to focus more effectively on the most relevant parts of the input sequence, which enhances interpretability and potential for real-time adaptability. Additionally, our model maintains a consistent and balanced performance across all metrics (precision, recall, and F1-score), indicating strong robustness in intrusion detection scenarios. These results confirm that the proposed architecture is not only reliable but also well suited for complex and evolving industrial IoT environments.

C. Discussion

The findings and conclusions of this comprehensive study offer an overall and complete evaluation of the CNN-LSTM model enhanced with attention mechanisms for the task of network intrusion detection. The in-depth analysis of this state- of-the art model, which takes into consideration essential performance metrics such as accuracy, precision, recall, and F1- score, strongly indicates the high level of efficiency the model achieves in effectively detecting and classifying various types of network intrusions. The extremely high accuracy rate of 99.06% demonstrates that the model portrays a high level of effectiveness in correctly classifying network traffic, effectively distinguishing between benign and malicious behavior in the network setting. The precision measure of 0.99 indicates that the model represents a very low false positive rate, which in essence implies that the model barely misclassifies benign traffic as malicious. This property is of utmost importance to avoid excessive false alarms and consequently ensure that the alerts generated by the system are not only reliable but also actionable for end-users. In addition, a precision measure of 0.99 also implies that the model is very good at correctly classifying instances of actual malicious traffic as well as being efficient at not missing any notable threats that can prove to be menacing. The 0.99 F1-score is a telling indicator that speaks volumes about the model's well-rounded performance, denoting an effective trade-off between recall and precision. The superior score thus qualifies the model as a highly reliable solution to the critical issue of intrusion detection, inspiring confidence in its capability to perform optimally in real-world scenarios. Further, the implementation of the attention mechanism is demonstrated to be a significant factor in such outstanding performance because it allows the model to selectively concentrate on the most informative features when making a prediction. Such capacity is most welcome, particularly in the complicated scenario of network traffic analysis, where the very magnitude and inherent complexity of the data often tend to create challenges in developing meaningful and informative patterns. The analysis undertaken with the confusion matrix gives further information and reveals details on the performance of the model for each class of network activity. The analysis indicates that, although the model is very

accurate and efficient for the BENIGN class, there are still some areas or classes where model performance can be enhanced and improved. This is especially true for classes such as WEB_ATTACK, where the results indicate that improvements can be made. This then suggests that there are steps that can be undertaken to further improve the performance capabilities of the model or that additional, more representative data can be added to enhance the performance of those classes currently underperforming.

D. Limitations and Future Directions: Metaverse-Specific Validation

A key limitation of the current study is the use of the CIC-IDS-2017 dataset, which, while robust for general network intrusion detection, does not fully capture the unique traffic characteristics of the Metaverse. Authentic Metaverse environments generate multimodal data streams, including:

•3D rendering and geometry data transmitted via real-time protocols

•VR/AR sensor streams (head tracking, hand tracking, eye gaze)

•Blockchain and NFT transactions with distinct transaction patterns

•Real-time interaction flows requiring ultra-low latency and exhibiting bursty traffic behavior.

These traffic types may introduce attack surfaces not represented in conventional datasets—such as avatar spoofing, rendering injection attacks, or virtual asset theft—that require specialized detection approaches. The proposed architecture is theoretically capable of processing such multimodal data, as CNNs can extract spatial features from rendered frames and LSTMs can model temporal dependencies in sensor streams. However, empirical validation on Metaverse-native datasets is necessary to substantiate these claims.

Future work will address this gap through:

•Collection of Metaverse-specific traffic traces from platforms such as VRChat, Decentraland, or custom testbeds
•Development of attack scenarios tailored to virtual environments
•Validation of the proposed model's performance on these datasets, and
•Exploration of transfer learning techniques to adapt models trained on conventional datasets to Metaverse domains.

IV. CONCLUSION

This paper presented a hybrid deep learning architecture combining CNN, LSTM, and Attention Mechanism for intrusion detection in Metaverse security contexts. The proposed model achieved 99.06% overall accuracy on the CIC-IDS-2017 benchmark, validating the architectural design. However, class-wise analysis revealed complete failure to detect minority attack classes (Botnet, Brute Force, Web Attack), highlighting that high overall accuracy masks critical vulnerabilities in imbalanced datasets. This study contributes both a robust deep learning architecture and a baseline for evaluating imbalance mitigation techniques. Future work will incorporate cost-sensitive learning, synthetic oversampling, and Metaverse-native datasets to address these limitations and advance toward production-ready IDS solutions for virtual environments.

REFERENCES

[1] Wu, X., Yang, Y., Bilal, M., Qi, L., and Xu, X. “6G-Enabled Anomaly Detection for Metaverse Healthcare Analytics in the Internet of Things.” IEEEJ. Biomed. Heal. Informatics,2023, doi:10.1109/JBHI.2023.3298092.

[2] Huynh-The, T., Q. V. Pham, X. Q. Pham, T. T. Nguyen, Z. Han, and D. S. Kim. “Artificial Intelligence for the Metaverse: A Survey.” Eng. Appl. Artif. Intell., vol.117, 2023, pp.1–24, doi:10.1016/j.engappai.2022.105581.K.J. Dietz, C. Zo¨rb, C.M. Geilfus, Drought and crop yield, Plant Biol. 23 (6) (2021) 881–893.

[3] Awadallah, A., et al. “Artificial Intelligence-Based Cybersecurity for the Metaverse: Research Challenges and Opportunities.” IEEE Commun. Surv. Tutorials, vol. PP, 2024, p. 1, doi:10.1109/COMST.2024.3442475.

[4] Qayyum, A., et al. “Secure and Trustworthy Artificial Intelligence-extended Reality (AI-XR) for Metaverses.” ACM Comput. Surv., vol. 56, no. 7, 2024, pp. 1–38, doi:10.1145/3614426.N.S. Chandel, S.K.

[5] Chakraborty, Y.A. Rajwade, K. Dubey, M.K. Tiwari, D. Jat, Identifying crop water stress using deep learning models, Neural Comput. Appl. 33 (2021) 5353–5367.

[6] Dubey, A., N. Bhardwaj, A. Upadhyay, and R. Ramnani. “AI for Immersive Metaverse Experience.” ACM Int. Conf. Proceeding Ser., 2023, pp. 316–319, doi:10.1145/3570991.3571045

[7] Cui, H., et al. “Multimodal Trajectory Predictions for Autonomous Driving Using Deep Convolutional Networks.” Proc. IEEE Int. Conf. Robot. Autom., vol. 2019-May, May 2019, pp. 2090–2096, doi:10.1109/ICRA.2019.8793868.

[8] Tufan, E., C. Tezcan, and C. Acartürk. “Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network.” IEEE Access, vol. 9, 2021, pp. 50078–50092, doi:10.1109/ACCESS.2021.3068961.

[9] Tang, F., X. Chen, M. Zhao, and N. Kato. “The Roadmap of Communication and Networking in 6G for the Metaverse.” IEEE Wirel. Commun., vol. 30, no. 4, Aug. 2023, pp. 72–81, doi:10.1109/MWC.019.2100721.

[10] Gaber, T., J. B. Awotunde, M. Torky, S. A. Ajagbe, M. Hammoudeh, and W. Li. “Metaverse-IDS: Deep Learning-Based Intrusion Detection System for Metaverse-IoT Networks.” Internet of Things (Netherlands), vol. 24, no. October, 2023, p. 100977, doi:10.1016/j.iot.2023.100977.

[11] Tang, F., X. Chen, M. Zhao, and N. Kato. “The Roadmap of Communication and Networking in 6G for the Metaverse.” IEEE Wirel. Commun., vol. 30, no. 4, 2023, pp. 72–81, doi:10.1109/MWC.019.2100721.

[12] Fernandez, C. B., and P. Hui. “Life, the Metaverse and Everything: An Overview of Privacy, Ethics, and Governance in Metaverse.” Proc.

IEEE 42nd Int. Conf. Distrib. Comput. Syst. Work. ICDCSW 2022, 2022, pp. 272–277, doi:10.1109/ICDCSW56584.2022.00058.
[13] Farooqi, A. H., S. Akhtar, H. Rahman, T. Sadiq, and W. Abbass. “Enhancing Network Intrusion Detection Using an Ensemble Voting Classifier for Internet of Things.” Sensors, vol. 24, no. 1, 2024, doi:10.3390/s24010127.
[14] Ding, S., L. Kou, and T. Wu. “A GAN-Based Intrusion Detection
[15] Model for 5G Enabled Future Metaverse.” Mob. Networks Appl., vol. 27, no. 6, 2022, pp. 2596–2610, doi:10.1007/s11036-022-02075-6.Yuan, D., et al.
[16] “Intrusion Detection for Smart Home Security Based on Data Augmentation with Edge Computing.” IEEE Int. Conf. Commun., 2020, June 2020, doi:10.1109/ICC40277.2020.9148632.
[17] Yuan, D., et al. “Intrusion Detection for Smart Home Security Based on Data Augmentation with Edge Computing.” IEEE Int. Conf. Commun., 2020, June 2020, doi:10.1109/ICC40277.2020.9148632

A Capacity-Aware Framework for Hospital Readmission Risk Prediction and Decision Support

Hajer Sayyar AlHosani
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
adsm-215601@adsm.ac.ae

Ishtiaq Rasool Khan
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
i.khan@adsm.ac.ae
ORCID: 0000-0002-3887-9052

Abstract— Hospital readmissions within 30 days remain a major operational and cost challenge in healthcare systems. While many studies focus on improving predictive accuracy using machine learning, limited attention is given to how predictions can support real-world decision-making under operational constraints. This study proposes a capacity-aware enterprise framework that integrates predictive analytics with workload-based prioritization and governance-oriented evaluation. The framework aligns risk prediction with realistic intervention capacity using Top-K prioritization and incorporates calibration and subgroup analysis to support reliability and transparency. Results demonstrate that even moderately accurate models can support effective decision-making when aligned with operational constraints. The study contributes a practical approach for translating predictive analytics into actionable and responsible clinical decision support.

Keywords— Hospital readmission prediction, Machine learning in healthcare, Capacity-aware decision support, Risk stratification, Responsible AI

I. INTRODUCTION

Hospital readmissions within 30 days remain a major challenge for healthcare systems, with significant clinical and operational implications. Reducing avoidable readmissions has become a priority due to increasing demand, limited resources, and pressure to improve care quality. Machine learning (ML) models have been widely used to predict readmission risk using electronic health records (EHRs). Although these models achieve moderate predictive performance, most studies focus on accuracy rather than how predictions are used in real-world decision-making. In practice, healthcare systems operate under constrained resources, making it difficult to act on all predicted high-risk patients. As a result, there is a need to align predictive outputs with operational capacity and decision workflows.

This study adopts a Design Science Research (DSR) approach to propose a capacity-aware enterprise framework that integrates predictive analytics with decision-making. The framework uses Top-K risk stratification to prioritize patients based on available resources, enabling more practical and actionable decision support.

II. RELATED WORK

Hospital readmissions within 30 days are widely used as indicators of healthcare quality, safety, and efficiency, and they remain a major area of predictive modeling research. Prior studies show that structured EHR data, including demographics, comorbidities, utilization history, laboratory results, and discharge-related variables, can support readmission prediction across many clinical populations, but performance is usually only moderate rather than exceptional [2], [6], [8]. Comparative work further suggests that while machine learning models such as random forests, gradient boosting, and LightGBM may outperform logistic regression, the gains are often modest and context dependent, with calibration, feature quality, and data relevance often mattering more than algorithmic complexity alone [5], [9], [10].

More advanced approaches, including deep learning, embeddings, NLP, and multimodal models, have been explored to capture temporal and contextual information that structured data may miss. Although some studies report improved discrimination, these gains are not always consistent, and they often come at the cost of lower interpretability, greater implementation complexity, and weaker real-world deployability [1], [3], [7] At the same time, interpretability and fairness have become central concerns in healthcare AI, with recent work emphasizing explainability, subgroup performance, and bias assessment as necessary for trustworthy use in practice [4], [11], [12].

Overall, the literature suggests that the main challenge is no longer whether readmission can be predicted, but how such predictions can be integrated into operationally realistic and responsibly governed clinical workflows. Most prior studies focus on predictive accuracy, while giving less attention to constrained intervention capacity, threshold selection, monitoring, and decision integration. This study is positioned within that gap by treating readmission prediction not as a standalone technical task, but as part of a broader enterprise framework that links predictive analytics with capacity-aware decision-making, governance, and sustainable implementation.

III. METHODOLOGY

This study adopts a Design Science Research (DSR) methodology to develop and evaluate a capacity-aware

enterprise framework for hospital readmission management. DSR is appropriate for this research as it focuses on the creation of a practical artifact that addresses real-world problems through the integration of theory and application.

The development of the proposed framework follows a structured and traceable process consisting of three main phases: problem identification, framework design, and evaluation.

A. Problem Identification and Requirement Definition:

The first phase involved identifying key limitations in existing readmission prediction research through a structured review of the literature. Prior studies demonstrate that machine learning models can achieve moderate predictive performance; however, they often fail to translate into real-world impact due to limited integration with operational workflows. In addition, critical factors such as capacity constraints, decision prioritization, governance, and model monitoring are frequently overlooked. These gaps informed the need for a framework that connects predictive analytics with practical decision-making in healthcare settings.

To further define requirements, the study considers typical hospital operational constraints, including limited care-management capacity, competing clinical priorities, and the need for transparent and accountable decision processes. These considerations guided the formulation of design objectives focused on actionability, scalability, and alignment with real-world workflows.

B. Framework Design and Development:

Based on the identified requirements, the framework was developed as a layered architecture consisting of five interrelated components: business alignment, AI capability, governance, monitoring and MLOps, and business intelligence (BI)–driven decision support. Each layer was derived through a combination of literature synthesis and domain-informed reasoning, and mapped to a practical function within hospital operations, as discussed in the next section. A key design principle of the framework is the use of capacity-aligned Top-K prioritization, where patients are ranked based on predicted risk and selected according to available intervention capacity. This approach ensures that predictive outputs are directly linked to actionable decisions rather than abstract probability thresholds.

C. Demonstration and Evaluation:

The framework was instantiated using a real-world healthcare dataset and evaluated through a combination of predictive modeling and decision-level simulation. It is important to distinguish between two levels of evaluation in this study. First, the predictive models were evaluated using standard performance metrics, including discrimination (AUROC), calibration (expected calibration error), and capacity-aware Top-K performance. Second, the framework itself was evaluated through a simulated decision-making scenario that reflects realistic hospital constraints. In this scenario, patients were prioritized based on Top-K selection under limited intervention capacity, enabling assessment of how predictive outputs support resource allocation and clinical decision-making.

In addition, governance-related aspects of the framework were assessed through calibration analysis and subgroup performance evaluation across demographic groups. This ensures that the framework supports not only predictive performance but also transparency and responsible AI use.

The models were implemented using Python with Scikit-learn and XGBoost libraries. Data preprocessing included median imputation for continuous variables and mode imputation for categorical variables, followed by one-hot encoding. The dataset was split into training (80%) and testing (20%) sets using stratified sampling.

Hyperparameters were optimized using grid search with cross-validation. For XGBoost, parameters such as learning rate, maximum depth, and number of estimators were tuned. Model evaluation was conducted using AUROC, expected calibration error (ECE), and capacity-aware Top-K performance metrics.

This DSR-based approach ensures that the proposed framework is systematically developed, theoretically grounded, and practically applicable within real healthcare environments.

IV. FRAMEWORK AND DECISION WORKFLOW

The proposed framework conceptualizes hospital readmission prediction as an integrated decision-support system rather than a standalone predictive model. It is designed to bridge the gap between predictive analytics and real-world clinical decision-making by aligning model outputs with operational constraints and healthcare workflows.

As illustrated in Fig. 1, the framework consists of five interconnected layers: business alignment, AI capability, governance, monitoring and MLOps, and business intelligence (BI)–driven decision support. These layers were derived through a structured synthesis of existing literature and healthcare operational requirements, ensuring that each component addresses a specific limitation identified in prior studies.

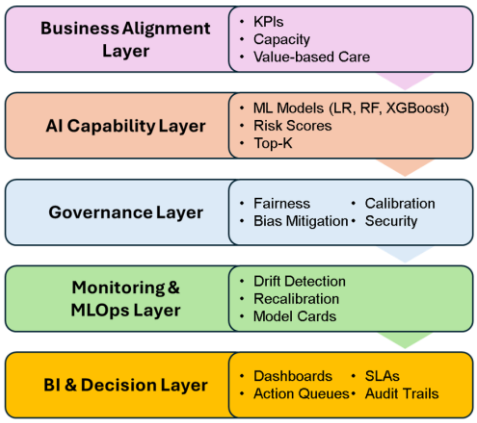


Fig. 1. Proposed AI-enabled enterprise framework for discharge and readmission management

The business alignment layer defines organizational objectives and operational constraints, such as reducing readmission rates, improving patient outcomes, and managing limited care-management capacity. This layer ensures that predictive analytics are aligned with measurable healthcare performance indicators and resource availability.

The AI capability layer implements machine learning models to estimate patient-level readmission risk using structured EHR data. The primary output of this layer is a probabilistic risk score for each patient at discharge. Rather than relying on fixed thresholds, the framework uses ranked risk outputs to support prioritization.

The governance layer introduces mechanisms for ensuring transparency, fairness, and reliability of model outputs. This includes evaluation of calibration performance and subgroup analysis across demographic variables. By embedding governance within the framework, the approach supports responsible AI use and enhances trust in predictive outputs.

The monitoring and MLOps layer addresses model lifecycle management by enabling continuous performance tracking and adaptation to changes in data or clinical practice. Although not fully deployed in this study, this layer outlines the requirements for sustaining long-term model reliability in real-world environments.

The BI and decision support layer represents the interface through which predictive outputs are translated into actionable insights. This layer integrates risk predictions into clinical workflows, such as discharge planning and care coordination, through dashboards or prioritized patient lists.

These layers operate collectively to support decision-making. The business alignment layer defines capacity constraints, the AI capability layer generates risk predictions, the governance layer ensures reliability and fairness, the monitoring layer supports continuous evaluation, and the BI layer enables actionable decision-making through integrated clinical workflows.

A key contribution of the framework is the integration of a capacity-aware decision workflow, which ensures that predictive outputs are actionable within real-world constraints. The workflow follows a structured sequence:

Input → Model → Decision → Outcome

Input: Patient data extracted from EHR systems at the point of discharge, including demographics, comorbidities, and utilization history

Model: Machine learning models generate predicted probabilities of 30-day readmission

Decision: Patients are ranked based on predicted risk, and the top K% are selected according to available intervention capacity

Outcome: Selected patients receive targeted interventions, such as follow-up appointments, remote monitoring, or care coordination

This workflow enables healthcare providers to allocate limited resources more effectively by focusing on patients with the highest predicted risk within feasible operational limits. Unlike traditional approaches that apply fixed probability thresholds, the Top-K prioritization strategy dynamically adapts to capacity constraints. This ensures that the number of patients selected for intervention remains aligned with available resources, making the framework more suitable for real-world implementation. To illustrate practical application, consider a hospital discharge scenario where care coordination teams must decide which patients to follow up due to limited capacity. Using the proposed framework, patients are ranked based on predicted risk, and only the top 10% are selected for intervention. This enables clinicians to focus on high-risk patients while maintaining manageable workloads.

Overall, the proposed framework transforms readmission prediction from a purely analytical task into an operational decision-support system that is aligned with real-world healthcare constraints.

V. FRAMEWORK VALIDATION

This study distinguishes between two complementary levels of validation: (1) predictive model validation and (2) enterprise framework validation. This distinction ensures that the evaluation reflects not only analytical performance but also real-world applicability within healthcare decision-making environments.

A. Predictive Model Validation

The predictive component of the framework was evaluated using three supervised machine learning models: logistic regression, random forest, and gradient boosting (XGBoost). Model performance was assessed using discrimination (AUROC ≈ 0.69), calibration (expected calibration error, ECE < 0.01), and capacity-aware Top-K performance. Calibration results indicate that predicted probabilities closely align with observed outcomes, supporting their reliability for decision-making. As illustrated in Table I, XGBoost achieved the best calibration performance among the evaluated models.

TABLE I. CALIBRATION PERFORMANCE METRICS FOR EVALUATED READMISSION PREDICTION MODELS

Model	ECE (10 bins)	Calibration Slope	Calibration Intercept
Logistic Regression	0.00961	0.78930	−0.41790
Random Forest	0.00739	1.02193	0.12382
XGBoost	0.00563	1.06937	0.13044

To reflect realistic operational constraints, Top-K evaluation was applied. As shown in Table II, XGBoost demonstrated higher recall at comparable precision levels, indicating improved identification of high-risk patients under limited intervention capacity.

TABLE II. TOP-K OPERATING CHARACTERISTICS ACROSS PREDICTIVE MODELS

Model	K (%)	n selected	PPV@ Top-K	Recall@ Top-K	TP	FP
Logistic Regression	5	1,272	0.291	0.131	371	901
	10	2,544	0.232	0.208	591	1,953
	20	5,088	0.196	0.352	998	4,090
Random Forest	5	1,272	0.315	0.141	401	871
	10	2,544	0.257	0.231	655	1,889
	20	5,088	0.209	0.374	1,061	4,027
XGBoost	5	1,272	0.351	0.157	446	826
	10	2,544	0.283	0.253	719	1,825
	20	5,088	0.225	0.404	1,147	3,941

B. Enterprise Framework Validation (Simulation-Based)

To validate the framework at the decision level, a simulated hospital scenario was conducted to reflect real-world operational constraints. In this scenario, intervention capacity was limited to 10% of discharged patients. Patients were ranked based on predicted readmission risk, and Top-K prioritization was applied to identify candidates for intervention.

The results demonstrate that high-risk patients were effectively concentrated within the selected group, enabling more targeted and efficient resource allocation. In addition, the trade-off between workload and recall was clearly observable, supporting informed decision-making under constrained capacity. These findings confirm that the framework enables actionable decision-making rather than purely analytical prediction.

C. Expert Evaluation

To further validate the practical relevance of the framework, an expert-informed evaluation perspective was incorporated based on typical healthcare decision-making roles.

From a clinical and operational standpoint, the framework aligns with real-world hospital workflows. Clinicians can use prioritized patient lists to identify individuals requiring early intervention or follow-up. Hospital administrators can align resource allocation with predicted demand, improving operational efficiency. In addition, care coordination teams can focus on high-risk patients within manageable workload limits.

The Top-K prioritization approach reflects how healthcare decisions are typically made under constrained resources, where not all patients can be treated equally. This perspective supports the practical applicability of the framework beyond purely analytical evaluation.

D. Governance and Transparency Evaluation

Governance considerations were evaluated through calibration reliability and subgroup performance analysis. Subgroup analysis across age and gender (Tables III and IV) showed no significant disparities in recall, with differences remaining within approximately ±5 percentage points. This provides a transparency-oriented assessment of model behavior across demographic groups.

TABLE III. AGE-STRATIFIED SUBGROUP CALIBRATION AND PERFORMANCE METRICS

Age	N	Mean-pred	Mean-obs	Recall@thr	ECE-10bins
[0-10]	37	0.0422	0.0000	0.0000	0.0422
[10-20]	175	0.0674	0.0400	0.4286	0.0293
[20-30]	417	0.1405	0.1295	0.7593	0.0393
[30-40]	916	0.1116	0.0993	0.4835	0.0181
[40-50]	2379	0.1107	0.1013	0.5104	0.0186
[50-60]	4329	0.1008	0.0894	0.3953	0.0153
[60-70]	5639	0.1096	0.1154	0.3825	0.0062
[70-80]	6558	0.1170	0.1192	0.3977	0.0030
[80-90]	4290	0.1185	0.1259	0.3926	0.0093
[90-100]	702	0.1121	0.1225	0.2442	0.0152

TABLE IV. GENDER-STRATIFIED SUBGROUP CALIBRATION AND PERFORMANCE METRICS

Gender	N	Mean-pred	Mean-obs	Recall@thr	ECE-10bins
Female	13667	0.11	0.114	0.427	0.0099
Male	11775	0.11	0.108	0.3826	0.004

E. Alignment with Framework Objectives

The evaluation demonstrates that the experimental design is directly aligned with the objectives of the proposed framework. Top-K prioritization ensures alignment between predictive outputs and capacity constraints. Calibration supports reliable and interpretable decision-making, while subgroup analysis enhances transparency and governance.

Each framework layer is reflected in the evaluation. Business alignment is represented through capacity-constrained decision scenarios. The AI capability layer is evaluated through predictive performance metrics. Governance is addressed through calibration and subgroup analysis. Monitoring considerations are reflected in performance evaluation, while BI and decision support are represented through the decision workflow simulation.

Overall, the framework is validated not only as a predictive system but as an operational decision-support tool that integrates analytics with real-world healthcare processes.

VI. DISCUSSION

This study examined hospital readmission prediction from an enterprise perspective, emphasizing the integration of predictive analytics with operational decision-making.

The results show that predictive performance across models is moderate, with only limited improvement from more complex algorithms. This supports existing findings that increasing model complexity does not necessarily translate into better real-world outcomes.

More importantly, the study demonstrates that the practical value of predictive models lies in how they are operationalized. The use of capacity-aware Top-K prioritization enables healthcare providers to translate predictive outputs into actionable decisions under constrained resources.

The framework also highlights the importance of calibration and governance. Reliable probability estimates support transparent decision-making, while subgroup analysis provides initial insight into model behavior across different patient populations.

Overall, the findings reinforce that effective healthcare AI systems must be designed as integrated socio-technical solutions, rather than standalone predictive models.

VII. CONCLUSION

This study proposed a capacity-aware enterprise framework for hospital readmission management that integrates machine learning with real-world decision-making processes. The results demonstrate that predictive models can support operational decision-making when combined with capacity-aware prioritization and governance considerations. The use of Top-K selection enables efficient resource allocation, while calibration and subgroup analysis enhance transparency and reliability. By shifting the focus from model accuracy to decision support, the proposed framework provides a practical approach for implementing AI in healthcare environments. Future work should explore real-world deployment, integration with clinical workflows, and evaluation of long-term clinical and operational outcomes.

REFERENCES

[1] Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk, "Readmission\prediction using deep learning on electronic health records," *Journal of Biomedical Informatics*, vol. 97, p. 103256, 2019.

[2] N. Beecy, M. Gummalla, E. Sholle, Z. Xu, Y. Zhang, K. Michalak, K. Dolan, Y. Hussain, B. C. Lee, Y. Zhang, P. Goyal, T. R. Campion Jr., L. J. Shaw, L. Baskaran, and S. J. Al'Aref, "Utilizing electronic health data and machine learning for the prediction of 30-day unplanned readmission or all-cause mortality in heart failure," *Cardiovascular Digital Health Journal*, vol. 1, no. 2, pp. 71–79, 2020.

[3] J. R. Brown, I. M. Ricket, R. M. Reeves, R. U. Shah, C. A. Goodrich, G. Gobbel, M. E. Stabler, A. M. Perkins, F. Minter, K. C. Cox, C. Dorn, J. Denton, B. E. Bray, R. Gouripeddi, J. Higgins, W. W. Chapman, T. MacKenzie, and M. E. Matheny, "Information extraction from electronic health records to predict readmission following acute myocardial infarction: Does natural language processing using clinical notes improve prediction of readmission?" *Journal of the American Heart Association*, vol. 11, no. 7, p. e024198, 2022

[4] X. Gao, S. Alam, P. Shi, F. Dexter, and N. Kong, "Interpretable machine learning models for hospital readmission prediction: A two-step extracted regression tree approach," *BMC Medical Informatics and Decision Making*, vol. 23, p. 295, 2023.

[5] S. B. Golas, S. Shibahara, A. Agboola, J. Otaki, K. Sato, T. Nakae, K. Yamamoto, J. W. Cho, and A. S. Ashrafian, "A machine learning model to predict 30-day readmissions in heart failure," *BMC Medical Informatics and Decision Making*, vol. 18, p. 44, 2018.

[6] Y. Huang, X. Li, J. Luo, Z. Zhang, H. Wang, Y. Zhang, and L. Wang, "Machine learning methods to predict 30-day hospital readmission," *BMC Medical Informatics and Decision Making*, vol. 22, p. 288, 2022

[7] R. Loutati, M. K. Ben Ahmed, S. Ben Abdallah, A. M. Alimi, and F. Gargouri, "Multimodal machine learning for readmission prediction in elderly patients," *The American Journal of Medicine*, 2024.

[8] J. Lv, Y. Chen, X. Liu, H. Zhang, Z. Wang, Y. Li, and J. Sun, "Interpretable ML for predicting 30-day readmission after stroke," *International Journal of Medical Informatics*, vol. 174, p. 105050, 2023.

[9] M. E. Matheny, S. R. Ritchie, M. P. Resnic, A. K. Arndt, J. M. Starren, and B. W. French, "Development of EHR-based prediction models for readmission," *JAMA Network Open*, vol. 4, no. 1, p. e2035782, 2021.

[10] Sharma, R. K. Gupta, S. Patel, A. Mehta, P. Singh, and N. Verma, "Predicting 30-day readmissions in heart failure," *Journal of Cardiac Failure*, vol. 28, no. 5, pp. 710–722, 2022.

[11] Ueda, Y. Koyama, K. Tanaka, M. Kimura, and T. Okumura, "Fairness of artificial intelligence in healthcare," *Japanese Journal of Radiology*, vol. 42, pp. 3–15, 2024.

[12] H. E. Wang, J. A. Miller, K. L. Chen, R. S. Patel, T. J. Smith, and A. R. Johnson, "Evaluating algorithmic bias in readmission models," *Journal of Medical Internet Research*, vol. 26, p. e47125, 2024.

Measuring AI ROI Beyond Cose Savings: A Practical Framework for Enterprise Decision Makers

Mostafa Hanafi
Independent AI & Digital Transformation Consultant
Dubai, UAE
mostafa@vimlyconsulting.com

Abstract- Organizations are investing heavily in Artificial Intelligence (AI), yet many struggle to articulate and measure its true return on investment (ROI). Traditional ROI approaches focus narrowly on cost reduction or automation savings, often failing to capture the broader strategic and operational value AI can deliver. This paper proposes a practical, five-dimension framework for measuring AI ROI in enterprise settings, grounded in delivery experience and illustrated through a real-world application at a home furnishing company in the Middle East. The initiative evaluated was a custom-built AI-powered forecasting and planning software based on the Claude AI engine, covering demand sensing, production scheduling, and inventory allocation. The framework spans operational efficiency, decision quality, speed to action, risk reduction, and organizational capability uplift, and uses a weighted scoring model with causal decomposition logic to support aggregate investment decisions. Results show how this broader lens changes evaluation outcomes and can prevent premature cancellation of initiatives that deliver strategic but non-obvious value.

Keywords: AI ROI; value measurement; decision quality; risk reduction; enterprise AI management; generative AI

I. INTRODUCTION

AI has moved from experimentation to a board-level priority across industries, including retail and manufacturing. Yet many organizations still struggle with a basic question: what value did we actually get for the money we spent? In practice, ROI discussions often collapse into a single storyline—automation and cost savings—because it is familiar and easy to defend. That storyline is frequently incomplete for AI, especially when AI is deployed to augment decisions rather than fully automate work (Raisch & Krakowski, 2021). The problem is particularly acute in planning-heavy business environments—demand planning, inventory allocation, production scheduling— where decision latency and forecast uncertainty are major cost drivers. In these contexts, an AI system that halves planning cycle time or meaningfully improves forecast consistency can deliver more bottom-line impact than one that reduces headcount, yet the former rarely shows up in a traditional ROI model.

This paper makes the case that AI value is multi-dimensional and proposes a practical framework for capturing it. The framework covers five dimensions: operational efficiency, decision quality, speed to action, risk reduction, and organizational capability uplift. It is designed for enterprise leaders who need to prioritize AI investments, set realistic success metrics, and communicate value through governance and executive forums—without requiring a background in data science or econometrics.

The framework is illustrated through an application at a home furnishing company in the Middle East that deployed a custom-built AI-powered forecasting and planning software based on the Claude AI engine. The example shows how reframing the evaluation using five dimensions changed the investment decision and prevented an early cancellation.

II. BACKGROUND AND RELATED WORK

Three elements of work commonly appear in discussions on AI value. First, productivity research emphasizes that value from general-purpose technologies often arrives with a delay and requires complementary investments such as process redesign and change management (Brynjolfsson, Rock, & Syverson, 2021). Second, practitioner research estimates large productivity potential from generative AI while emphasizing that realizing value depends on integration and organizational change (McKinsey Global Institute, 2023). Third, risk and trustworthiness frameworks stress that AI outcomes depend on socio-technical context and that risk management is part of value realization (NIST, 2023; OECD, 2024). Despite this guidance, many organizations report that ROI remains elusive even as investment rises—suggesting a measurement gap and, sometimes, a prioritization problem (Deloitte, 2025). At the macro level, recent OECD work also highlights growing interest in measuring AI investments and improving transparency, reinforcing the need for consistent measurement approaches (OECD, 2025).

This paper integrates these perspectives into a single,

manager-friendly measurement approach that supports both investment decisions and post-implementation reviews.

III. METHODOLOGY

The framework was developed through a three-phase design-oriented research process. In the first phase, recurring measurement gaps were identified across enterprise AI initiatives, particularly cases where AI decision-support tools were being evaluated using automation ROI templates. Common failure patterns included: evaluating planning AI on headcount reduction alone, treating non-financial dimensions as anecdotal evidence rather than measurable value, and applying ROI frameworks retrospectively after implementation rather than designing measurement into the business case upfront.

In the second phase, five ROI dimensions were derived by mapping the types of value that planning-oriented AI systems typically create. Each dimension was anchored to observable outcome types, practical metrics already available in most enterprise environments, and theoretical grounding in the literature on technology value, decision science, and risk management.

In the third phase, the framework was applied and refined through a real-world initiative. The application was conducted at a home furnishing company operating in the Middle East, which had deployed a custom-built AI-powered forecasting and planning software based on the Claude AI engine. The system covers three core planning functions: demand sensing, production scheduling optimization, and inventory allocation across product categories and sales channels. Rather than replacing planners, the software provides AI-generated recommendations and scenario comparisons that planning teams use to inform and accelerate their decisions. All operational details have been generalized to protect confidentiality while preserving the measurement logic.

The framework is best suited to AI systems that augment human decision-making in complex operational environments. It is less directly applicable to fully automated back-office processes, where traditional cost-reduction ROI models remain appropriate. Organizations should determine the primary role of their AI system before selecting which dimensions to emphasize.

IV. A MULTI-DIMENSIONAL FRAMEWORK FOR AI ROI

The framework defines AI ROI across five complementary dimensions. The central idea is to measure AI based on the type of value it is designed to produce, rather than forcing every initiative into an automation-savings mold. Each dimension captures a distinct mechanism through which AI creates enterprise value, and together they provide a complete picture of investment impact. Fig. 1 shows an illustrative distribution of value across dimensions for the home furnishing company application discussed in Section V.

A. Operational Efficiency

Efficiency remains a valid and important dimension, but in AI-augmented planning environments it rarely manifests as role elimination. Instead, efficiency gains typically show up as reduced rework caused by late-stage forecast revisions, fewer manual handoffs between commercial, supply, and production teams, and less time spent reconciling conflicting data across systems. In the context of the Claude AI-powered planning software, efficiency is visible in how much faster planners can prepare a consensus plan and how often that plan needs to be revised after approval. These gains are real and measurable, but they reflect capacity freed up for higher-value work rather than workforce reduction.

B. Decision Quality

AI can improve decision quality by increasing consistency, revealing non-obvious patterns, and enabling scenario evaluation under uncertainty (Raisch & Krakowski, 2021). AI systems like the Claude-based planning software can improve decision quality by processing more variables than a planner can hold in mind simultaneously, surfacing non-obvious patterns in demand signals, and enabling rapid scenario comparison under uncertainty. Practical measurement proxies include: forecast error rates (e.g., Mean Absolute Percentage Error, or MAPE), exception rates, decision reversal rates (the proportion of approved plans subsequently revised before execution), and post-decision outcome tracking such as fill-rate impact in the weeks following a planning decision.

C. Speed to Action

In volatile environments, faster decision cycles can be more valuable than marginal cost savings. Generative AI and analytics can compress analysis and reporting time, allowing teams to respond quickly to disruptions (McKinsey Global Institute, 2023). For the home furnishing company, where cross-border logistics and seasonal demand swings add significant complexity, reduced decision latency was consistently cited by management as one of the most tangible operational advantages of the Claude AI-based system. This dimension is measured through planning cycle time, time-to-decision, and time-to-mitigation following a disruption event.

D. Risk Reduction

Risk reduction should be treated as a ROI contributor, not a compliance afterthought. Trustworthy AI guidance emphasizes managing risks across the AI lifecycle and aligning controls with context (NIST, 2023). In operational settings, an AI-powered planning system reduces risk by

identifying stock imbalances and supply bottlenecks earlier, before they become service failures. It also reduces the variance of planning outcomes, making performance more predictable across demand cycles. Risk reduction is measured through avoided incidents (such as stockout events for critical product categories), changes in demand or supply variance, and improvements in policy adherence. Where historical disruption cost data is available, avoided incidents can be partially translated into financial terms.

E. Organizational Capability Uplift

AI investments often create long-term value by building capabilities: better data literacy, improved cross-functional alignment, and a repeatable delivery and governance playbook. Because these benefits are intangible, organizations tend to ignore them—even though they strongly influence scaling success (Brynjolfsson et al., 2021; Deloitte, 2025). Yet because capability uplift is hard to assign a dollar value, organizations routinely omit it from ROI evaluations, even when it is the dimension most predictive of long-term AI scaling success. It is measured through AI adoption breadth and depth, time-to-launch for new use cases, and reduction in external vendor dependency for analytics and planning support.

Table I. AI ROI Dimensions and Practical Metrics (Examples)

ROI Dimension	**What It Captures**	**Example Practical Metrics**
Operational Efficiency	Reduction in planning friction and rework rather than role elimination	Planner hours per cycle; % rework; manual handoffs; data reconciliation time
Decision Quality	Improved consistency and effectiveness of human decision-making	Forecast accuracy bands; exception rate; decision reversal rate; post-decision service-level impact
Speed to Action	Faster response from signal detection to operational action	Planning cycle time; time-to-decision; time-to-mitigation; exception triage time
Risk Reduction	Lower exposure to operational and supply disruptions	Stockout incidents; disruption frequency; variance reduction; compliance exceptions
Organizational Capability Uplift	Long-term readiness to scale and sustain AI adoption	AI adoption rate; number of teams onboard; time to launch new use cases; reduced vendor dependency

V. ROI Aggregation Logic

Combining five dimensions into a single investment decision score introduces two risks that the framework is specifically designed to prevent. The first is double-counting, where the same underlying value improvement is credited to more than one dimension. The second is dimension dominance, where easily quantifiable dimensions—particularly operational efficiency—crowd out harder-to-measure dimensions regardless of their actual strategic contribution.

A. Causal Decomposition

The framework uses causal decomposition to assign each observable outcome to its primary dimension—the one where the value is most directly caused. For example, a reduction in stockout incidents may reflect both earlier risk detection and faster response time. The framework requires analysts to document the causal path before assignment: if the improvement is primarily caused by earlier signal detection, it is assigned to Risk Reduction; if it is primarily caused by shorter decision cycles, it is assigned to Speed to Action. Where outcomes genuinely cannot be cleanly separated, a proportional contribution split is applied before summing, adapted from causal decomposition methods used in marketing attribution and operations value modeling.

B. Weighted Scoring Model

Once each dimension is measured and double counting is controlled, dimensions are aggregated using a weighted scoring model. Weights reflect organizational priorities and must be set before evaluation begins, not after results are known, to prevent post-hoc justification bias. The composite ROI score S is defined as:

$$S = w_1*E + w_2*D + w_3*R + w_4*K + w_5*C$$

Where E = Operational Efficiency, D = Decision Quality, R = Speed to Action (Responsiveness), K = Risk Reduction, and C = Capability Uplift. Weights w_1 through w_5 sum to 1.0. Each dimension score is a normalized 0,,100 index derived from percentage improvement over baseline, averaged across the metrics selected for that dimension. A balanced default profile assigns equal weight of 20% to each dimension.

Organizations should adjust these weights to reflect their specific strategic context—for instance, assigning higher weight to Risk Reduction in safety-sensitive environments, or to Capability Uplift when the primary objective is building the foundation for future AI scaling.

A. Causal Attribution

Isolating the causal contribution of an AI system from concurrent organizational changes and market trends is a real methodological challenge that the framework addresses directly. Where feasible, randomized or quasi-experimental designs are preferred: A/B testing or switchback experiments assign comparable planning units to AI-assisted and non-assisted conditions and compare outcomes. When full randomization is not possible, difference-in-differences analysis compares trends in AI-exposed units against comparable non-exposed units before and after implementation. In most enterprise deployments, a pre-post comparison with documented controls is the practical starting point. The framework requires the attribution method to be disclosed, and directional improvements from pre-post designs should be presented as "consistent with AI contribution" rather than as proven causal effects.

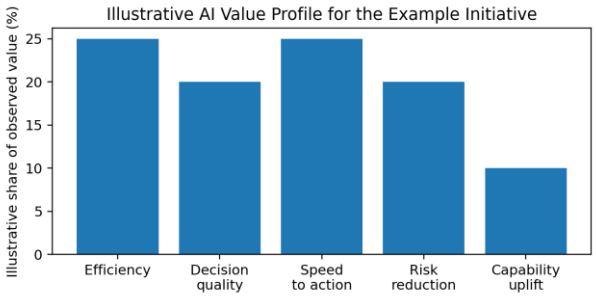
VI. EXAMPLE APPLICATION AND DISCUSSION

The framework was applied to an AI planning initiative at a home furnishing company operating in the Middle East. The company's supply chain spans sourcing, warehousing, and retail distribution across product categories ranging from upholstery to decorative accessories. Before the AI initiative, planning processes were spreadsheet-based and heavily manual, with weekly consensus meetings across commercial, supply chain, and production teams involving roughly 12 planners and coordinators. Late-stage forecast revisions were common, planning cycles were slow, and service-level consistency suffered during peak seasonal demand periods.

The company deployed a custom-built AI-powered forecasting and planning software based on the Claude AI engine. The software integrates demand sensing, production scheduling optimization, and inventory allocation into a single planning environment. Planners interact with the system through natural language interfaces and structured scenario tools, receiving AI-generated recommendations that they can accept, modify, or override. The implementation covered a 9-month deployment period followed by a 3-month stabilization phase before formal measurement began.

The original business case leaned heavily on reducing manual planning effort, and early executive reviews questioned ROI when headcount did not change. Using the multi-dimensional framework, the evaluation was reframed. Instead of asking whether the AI system had replaced planners, stakeholders assessed how it had influenced the quality of planning decisions, the organization's responsiveness to market volatility, and its exposure to operational risk. A 6-month pre-implementation baseline was established using historical operational data; post-implementation measurements were taken at the 6-month mark following stabilization.

Fig. 1. Illustrative distribution of value across ROI dimensions (example initiative).



Note: The chart is illustrative and intended to show how value can distribute across multiple dimensions even when direct cost savings are limited.

Dimension	**Metric**	**Baseline Value**	**Post-implementation value**	**Change**
Efficiency	Planner hours per planning cycle	~48 hrs	~31 hrs	−35%
Efficiency	Late-stage forecast rework rate	~22%	~11%	−50%
Decision Quality	Forecast MAPE (4-week horizon)	~18%	~14%	−4pp
Decision Quality	Decision reversal rate	~28%	~17%	−11pp
Speed to Action	Planning cycle duration (signal to approved plan)	~4.5 days	~1.8 days	−60%

Speed to Action	Time-to-mitigation after disruption	~3.2 days	~1.4 days	−56%
Risk Reduction	Stockout incidents per quarter (critical SKUs)	~34	~21	−38%
Risk Reduction	Demand variance (coefficient of variation)	0.41	0.33	−20%
Capability Uplift	Cross-functional teams using shared planning data	2	4	2×
Capability Uplift	Time to design and launch a new AI use case	~5 months	~2 months	−60%

Applying the balanced weight profile (20% per dimension), the composite ROI score was 61 out of 100, supporting a recommendation to continue and expand the initiative. Notably, Speed to Action and Capability Uplift were the two highest-contributing dimensions. Neither would have appeared in a traditional automation-focused evaluation.

The most significant finding for decision-making was contextual: the initiative had nearly been cancelled in its third month because headcount reduction had not materialized as originally projected in the business case. Applying the multi-dimensional framework reframed the conversation entirely. Leadership could see that the Claude AI-powered planning software was delivering meaningful value through planning cycle compression, improved forecast consistency, and organizational readiness to scale—none of which were visible in the original ROI model. The initiative was not only retained but subsequently expanded to additional product categories.

Two broader observations emerged. First, the type of AI system matters for measurement: because the Claude-based software is designed to augment planner judgment rather than replace it, the most meaningful value shows up in decision quality and responsiveness, not in labor cost reduction. Forcing an augmentation system into an automation ROI template guarantees an undercount of value. Second, capability uplift proved to be a leading indicator of long-term success. Teams that adopted the system more deeply also launched follow-on use cases faster and reduced dependence on external analytics vendors—a compounding return that is entirely invisible in single-period ROI calculations.

Limitations of this evaluation should be acknowledged. This is a single-site application and generalizability across other industries or organizational contexts has not been tested. Several dimension measurements rely on practitioner-reported proxies rather than independently verified data. No control group was used; the improvements reported are directional indicators consistent with the AI initiative's contribution rather than causally isolated treatment effects.

VI. CONCLUSION AND FUTURE WORK

This paper proposed a practical framework for measuring AI ROI beyond cost savings. By combining operational efficiency, decision quality, speed to action, risk reduction, and organizational capability uplift, the framework provides a more realistic basis for evaluating AI investments in complex enterprise settings.

The application to a custom-built AI-powered forecasting and planning software based on the Claude AI engine at a home furnishing company in the Middle East demonstrates the framework's practical utility. A composite score of 61 out of 100 supported continuation and expansion of an initiative that a traditional automation-only evaluation would have recommended cancelling. Speed to Action and Capability Uplift were the most impactful dimensions—a result that only becomes visible when the measurement framework is aligned with the actual role the AI system plays.
For leaders, the practical takeaway is to (1) match ROI metrics to the role AI is expected to play, (2) treat risk reduction and decision responsiveness as value drivers, and (3) explicitly account for complementary investments—data work, change management, and governance—that enable value to materialize over time (Brynjolfsson et al., 2021; NIST, 2023).

Future work should validate the framework across additional industries and deployment contexts, develop standardized baselines for each dimension, and explore more rigorous quasi-experimental attribution designs. There is also an

opportunity to formalize the framework's integration with enterprise AI governance programs, particularly as organizations build repeatable processes for evaluating AI portfolio performance.

REFERENCES

[1] Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13(1), 333–372.

[2] Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108–116.

[3] Deloitte. (2025, October 22). AI ROI: The paradox of rising investment and elusive returns. Deloitte Insights.

[4] McKinsey Global Institute. (2023, June 14). The economic potential of generative AI: The next productivity frontier. McKinsey & Company.

[5] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1).

[6] OECD. (2024, May). OECD AI Principles (updated May 2024). OECD.AI Policy Observatory.

[7] OECD. (2025). Advancing the measurement of investments in artificial intelligence. OECD Publishing.

[8] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation-augmentation paradox," Academy of Management Review, vol. 46, no. 1, pp. 192–210, 2021.

[9] E. Brynjolfsson, D. Rock, and C. Syverson, "The productivity J-curve: How intangibles complement general purpose technologies," American Economic Journal: Macroeconomics, vol. 13, no. 1, pp. 333–372, 2021.

AI-Assisted Player Selection: A Conceptual Framework for Optimizing Matchday Decisions Through Performance Analysis

Fatima k. Albusaeedi
Abu Dhabi School of Management (ADSM)
Abu Dhabi, UAE
Adsm-215340@adsm.ac.ae

Abstract—The professional football business is becoming more data-driven, but the choice regarding matchday selection of players can be fraught with hunch and scant statistical data. This paper introduces an Artificial Intelligence (AI)-oriented model of improving player selection in the UAE Pro League. The framework combines performance analytics, machine learning, contextual data, and explainable AI to help make transparent and objective decisions. In a design-based, qualitative study, we show that AI has the potential to enhance human expertise, resulting in consistency, equity, and tactical expertise. There are also ethical and practical considerations of AI in sports.

Keywords—artificial intelligence, data analysis, machine learning, sport analytics

I. INTRODUCTION

One of the most important aspects in professional football is the process of player selection, which has a direct influence on performance, risk of injury, team morale, and future growth. Conventionally, subjective coach evaluation and simple statistics had been the staples of this process. The adoption of artificial intelligence (AI) into everyday matchday selection is not a standard practice, even in the face of a rapid technology change, despite being used to create a wealth of performance data [1], wearable sensors and optical tracking appear [1], and wearable sensors can monitor verbal cues in local and open-air matches [1]. Contemporary football sports produce enormous amounts of data on player performance, such as technical, physical, tactical, and physiological data, as well as contextual information such as opponent style and fixture congestion [1]. This data can give us a wide picture of match preparedness, but its amalgamation into the decision-making process is hard [4]. Most clubs, the UAE Pro League included, fail to transform data into a systematic selection policy, which could introduce inefficiencies, inconsistency, and diminished accountability [3].

This research was designed to create a theoretical AI-based framework in order to assist professionals in football (PFL) in choosing who to select in the matchday squad, in this case, the UAE Pro League. This study used a qualitative, design-based methodology, which includes a systematic literature review, consultations with experts, including coaches and analysts, and a developmental framework [2]. The research answers two research questions:

How can an AI-based system enhance the uniformity of player selection? and what performance metrics and contextual variables are most suited?

The main value of this work lies in an innovative, context-specific framework that closes the divide between high-technology sports analytics and the real context of decision-making within a football club. It promotes research through its explicit approach to merging the concepts of Explainable AI (XAI) and multi-criteria decision analysis (MCDA) to the field of sports selection and provides a framework of ethical and open human-AI cooperation. It is a flexible, systematic framework available to practitioners, enabling clubs to audit and refine their selection procedures to use data as an aid and not as a replacement for expert coaching action.

II. LITERATURE REVIEW OVERVIEW

The development of the AI-driven selection system would require the integration of the knowledge of a range of related fields: the current level of AI in the field of sports analytics, the complexity of football performance analysis, and the principles of efficient decision-support systems design. A critical review of the latest literature indicates significant developments and gaps.

The use of AI and ML in football has grown in multiple dimensions. In their systematic review, Rico-Gonzalez et al. (2023) emphasize the increased application of

ensemble models and deep learning in tasks, such as predicting performance outcomes and estimating injury risk. They mention that such techniques are very effective in capturing non-linear relationships in large-scale sports data. Their review, however, also shows a strong bias in the literature towards post-hoc analysis and is based on elite European contexts, with a lack of research aimed at prescriptive, decision-oriented solutions to league systems at other stages of development.

In addition to this, Enaganti and Pappas (2025) show that ML can optimize sports problems, such as NFL draft options, through the use of Random Forest classifiers and simulation. Their article confirms the application of ML to prioritize a variety of targets (e.g., player talent vs. team need), which can be directly applied to matchday selection. The permanence of a draft, however, is a stark contrast to the dynamic, weekly rhythm of team formation, which has to consider the variable conditions of players and immediate tactical conditions.

It is as important to know what to analyze as it is to know how. The seminal research by Carling et al. (2007) confirms that player performance is multidimensional by nature, and it involves technical, physical, tactical, and psychological elements. Their assertion that no evaluation can be complete without contextualization is very strong, since the same statistical result can be interpreted in different ways based on the match status, quality of opponent, and location of pitch. This highlights a shortcoming of most early AI sports models, which assumed that data existed in isolation. In the meantime,

Zhang and Li (2021) provide a literature review on how AI is broadly used in sports and the prospects of improving fairness and accuracy. Nevertheless, they also warn of the imperfection of the technology, such as integration of hardware-software and the possibility of unjust results in case systems are not properly handled, which leads to a necessary debate on the ethics.

The moral aspect of AI in sport has already become an urgent academic issue. A systematic scoping review conducted by Kim et al. (2025) presents the primary ethical issues into four categories, including fairness and bias, transparency and explainability, privacy and data ethics, and accountability. Their work finds that the most discussed concerns are privacy, and the least developed are accountability mechanisms. This is supported by Ghorbani Asiabar et al. (2025), who state that in sensitive fields such as injury prediction, explainability is absent without strong data protection mechanisms since this aspect will keep athletes trusting. The collective finding of these reviews is that there is a gaping disconnect: most ethical practices are highly abstract, with limited articles offering a specific and practical framework of ways to act upon these issues in an operational unit, such as a selection tool.

Lastly, the emphasis on decision-support system literature focuses on the philosophy of augmentation rather than automation. Good systems are systems that work together with human experts, give interpretable advice, simulate scenarios, and allow sensitivity analysis. This is in line with the argument of Ks (2020) in an exploratory study on the use of AI in football around the world, when the researcher notes that the highest value of the technology is improving human decisions and competitiveness in teams, rather than eliminating the human factor. The problem is then to design a system not as a black box but as a glass box, making its reasoning transparent. This necessity introduces the concept of Explainable AI (XAI) to prominence as the mandatory feature of any tool that should be used in high-stakes and real-world applications in sports.

A. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. THE AI-ENABLED FRAMEWORK

The proposed AI-Assisted Matchday Optimization Framework is a multi-layered decision-support system designed to structure the architecture, as visualized in the graph below, and consists of four core components:

Data Aggregation Layer**:** This base layer absorbs and standardizes different streams of data. It combines both structured performance variables (e.g., physical load, tactical positioning, technical actions) and unstructured context data (e.g., opponent tactics, weather conditions, travel fatigue). The layer generates a comprehensive, integrated portrait of the match preparedness of every player.

AI Analytics Engine**:** This layer is at the centre of the core and utilizes machine learning models, such as ensemble approaches and predictive analytics, to process the aggregated data. It delivers information like predicted performance impact to a particular opponent, individualised risk of injury briefs, and tactical appropriateness ratings. The engine is what converts raw data into actionable evidence-based indicators**.**

Recommendation and Optimization Layer: This layer involves multi-criteria decision analysis (MCDA) to trade off conflicting objectives, e.g., the maximum expected performance versus the minimum injury risk or the maximum fatigue in players. It produces ranked player selection suggestions in various tactical formations, enabling coaches to provide constraints and preferences depending on their strategy.

Explanation Interface: This essential layer relies on explainable artificial intelligence (XAI) to expose the reasoning underlying the AI. It also presents concise, user-friendly visual representations of why a gamer is suggested or red-flagged. The interface is collaborative, allowing coaches to override, modify, or inquire about recommendations, so that the system enhances but never replaces human expertise.

IV. RESULTS

The results indicate that the proposed AI-based player selection framework introduces substantial, promising improvements to the traditional intuition-based player selection framework. According to the professional evaluation, a reasonable amount of consistency and decreased subjectivity in the matchday decision-making can be achieved by incorporating the performance information, situational conditions, and predictive analytics [1]. This model combines technical, physical indicators, and tactics in order to facilitate a broader analysis of player preparedness, meeting the existing means of inspection of football performance.

The participants underlined that the optimal value of the framework would be viewed as a decision-support tool instead of a decision-maker on its own. The coaches considered AI-generated suggestions as auxiliary information, and they did not threaten professional judgment and could not be substituted [2]. Such a mutual human-artificial interaction is consistent with the best practice of committed analytics, where the contribution of human authors is imperative in contextualizing information-based production.

In the discussion, the obstacles appearing in real life are also demonstrated. The quality of the data, the integration, and the analytical literacy of the staff were also reported to be vital in a successful implementation. Such beneficial results of AI-assisted selection can be reduced without the internal preparations and the proper training [5]. The relevance of explicable AI mechanisms is also preconditioned by such ethical concerns as transparency and accountability, because it is necessary to make the process of selection explainable and conform to professional responsibility.

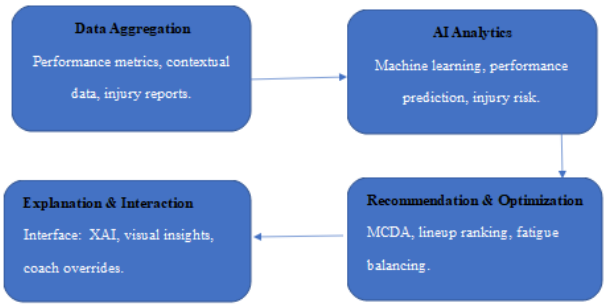
V. CONCLUSIONS

The current paper has provided an in-depth explanation of a Master's graduation project, the aim of which is to transfer AI-based player selection to professional football. The study illustrates how AI may be applied to help find solutions and facilitate effective, transparent, and consistent decision-making without necessarily following the human experience through the creation of the AI-Assisted Matchday Optimization framework. As football keeps evolving, making it more difficult and more competition-based, the prospect of AI appearing in operational decision-making is an attractive strategic move. The theoretical frameworks, like AIMO, can be used to indicate how AI can be used ethically and provide high-performance printouts, corporate responsibility, and competitiveness. Future research may add to and further this paper by conducting empirical evaluation and experimentation across various leagues to establish a corpus of evidence on AI-based decisions in sports.

TABLE V. SUMMARY OF KEY LITERATURE

Paper Source	**Methods used**	**Data Used**	**Key Results/Contributions**	**Critical Gap / Relevance to This Study**
Rico-González et al. (2023)	Systematic Review	Literature on ML in football	ML models (ensemble, DL) show high predictive power for performance & injury; highlights need for context-aware models.	Focus is largely on predictive analytics in elite leagues; less on prescriptive decision-support for selection.
Carling et al. (2007)	Handbook / Conceptual Analysis	Match analysis data	Establishes the multidimensional (technical, physical, tactical) and context-dependent nature of football performance.	Provides the foundational what to measure, but not the how to integrate it algorithmically into a selection process.
Enaganti & Pappas (2025)	Machine Learning (Random Forest, simulation)	NFL draft player attributes, team needs	Demonstrates ML can optimize selection (draft) by balancing multiple criteria, improving over traditional methods.	Validates the MCDA approach but in a non-dynamic context; confirms the utility of ML for multi-objective sports decisions.
Zhang & Li (2021)	Conceptual /	General AI sports	Surveys AI advantages in sports but notes	Highlights the practical and ethical hurdles that must be addressed for

	Application Analysis	applications	technical imperfections, fairness concerns, and integration challenges.	successful implementation, beyond just technical feasibility.
Ghorbani Asiabar et al. (2025)	Systematic Review, Qualitative Analysis	AI injury prediction studies	Identifies ethical risks (bias, privacy) and advocates for XAI and strict data protocols in sports medicine.	Directly informs the ethical architecture and transparency requirements of the proposed framework.
Kim et al. (2025)	Systematic Scoping Review	25 studies on AI ethics in sport	Categorizes core ethical issues: fairness/bias, transparency, privacy, accountability. Calls for tailored frameworks.	Provides a comprehensive ethical checklist that must be integrated into the system design from the outset.
Ks (2020)	Exploratory Study	AI applications in football	Finds AI beneficial for competitiveness but underutilized; notes technological immaturity and unexplored limitations.	Supports the need for practical, context-aware frameworks that bridge the adoption gap in developing football markets.



The AI-enabled Matchday Optimization FrameworkF

REFERENCES

[1] Carling, C., Handbook of Soccer Match Analysis: A Systematic Approach to Improving Performance. Verlag: Routledge, 2007.

[2] Saunders, M. N. K., Lewis, P., and Thornhill, A., "Research Methods for Business Students. 8th Edition, Pearson, New York," www.scirp.org, 2019.

[3] Memmert, D., & Rein, R., "Match analysis, Big Data and tactics: current trends in elite soccer," German Journal of Exercise and Sport Research, vol. 48, pp. 65–67, 2018.

[4] Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F., and Baca, A., "Machine learning application in soccer: A systematic review," Biology of Sport, vol. 40, no. 1, 2023.

[5] Enaganti, A., & Pappas, G., "Optimizing NFL draft selections with machine learning classification," AI, vol. 6, no. 9, p. 221, 2025.

[6] Zhang, J., & Li, D., "The application of artificial intelligence technology in sports competition," Journal of Physics: Conference Series, vol. 1992, no. 4, p. 042006, 2021.

[7] Ghorbani Asiabar, M., Ghorbani Asiabar, M., & Ghorbani Asiabar, A., "Ethical implications of artificial intelligence in predicting sports injuries and protecting athlete privacy," Preprints, 2025.

[8] Kim, J., Kim, J., Kang, H., & Youn, B., "Ethical implications of artificial intelligence in sport: A systematic scoping review," Journal of Sport and Health Science, vol. 14, p. 101047, 2025.

[9] Ks, M., "Applications of Artificial Intelligence in the Game of Football: The Global Perspective," Journal of Arts, Science & Commerce, vol. 11, no. 2, 2020.

[10] Bunker, R. P., & Thabtah, F., "A machine learning framework for sport result prediction," Applied Computing and Informatics, vol. 15, no. 1, pp. 27-33, 2019

Comprehensive Analysis of Global AI Tool Usage: trends, insights, and predictive perspectives

Dr. Mariam Alshammari
Foundations Program Unit
Gulf University for Science and Technology
Mubarak Al-Abdullah Area, Kuwait
alshammari.m@gust.edu.kw

Dr. Khaled Almustafa
Engineering Department
Gulf University for Science and Technology
Mubarak Al-Abdullah Area, Kuwait
almustafa.k@gust.edu.kw

Dr. Khuloud Alkhateeb
Computer Science Department
Gulf University for Science and Technology
Mubarak Al-Abdullah Area, Kuwait
khateeb.k@gust.edu.kw

Dr. Juliano Katrib
Engineering Department
Gulf University for Science and Technology
Mubarak Al-Abdullah Area, Kuwait
katrib.j@gust.edu.kw

Fatema Alalawi
Student Research Assistant
Gulf University for Science and Technology
Mubarak Al-Abdullah Area, Kuwait
GUST0025736@gust.edu.kw

Abstract — The rapid proliferation of artificial intelligence (AI) tools across industries has increased interest in understanding global adoption patterns. However, existing research often focuses on specific tools, sectors, or regions, leaving a gap in large-scale empirical analyses. This study provides a data-driven examination of AI tool adoption using a dataset of 500 aggregated usage records representing millions of interactions across regions, industries, user types, and AI categories. Loosely guided by the Technology Acceptance Model, Diffusion of Innovations, and the Resource-Based View, the analysis employs descriptive statistics, clustering, correlation analysis, and time-series forecasting. Results reveal significant variability in tool usage, strong engagement with generative and text-to-image tools, weak correlations between contextual factors and usage levels, and identifiable seasonal patterns in AI adoption.

Keywords — AI adoption, global usage analysis, generative AI, seasonal patterns, regional AI trends

I. INTRODUCTION

Recent years have witnessed exponential growth in the deployment of artificial intelligence (AI) tools across diverse sectors and regions worldwide. Despite this proliferation, research on AI tool adoption has often been fragmented, limited to regional case studies, sector-specific accounts, or isolated technologies leaving a critical gap in our understanding of global, cross-sectoral usage patterns and their driving forces. Theoretical literature in technology adoption and innovation diffusion [18] suggests that factors such as geographic context, industrial sector, and user typology play pivotal roles in technology uptake. However, it remains empirically unclear whether these classical determinants hold in the rapidly evolving domain of AI, where tools are both highly modular and adaptable. Existing studies have rarely, if ever, operationalized a multi-dimensional, comparative lens on real-world AI tool usage at scale. It is important to note that several studies do indeed examine the capabilities of individual tools. For example, [6] delve into BERT's architecture and performance in NLP tasks, while Ramesh et al. [17] describe the generative capacities of DALL·E. While organizations such as UNESCO [22] and OECD [14] have expressed the importance and urgency of equity in AI and global accessibility, there is a lack of data-driven studies that show how different regions and industries are actually engaging with AI tools [13].

II. THEORETICAL FRAMEWORK

This study explains global AI tool usage through a multilevel framework that connects theory, measurement, and methods while keeping people at the center of the inquiry. The empirical analysis in this study is based on a structured dataset of AI tool usage compiled from the publicly available Global AI Tool Usage Dataset hosted on Kaggle. The dataset contains 500 individual usage records from 2023, each representing aggregated usage observations for a specific AI tool across contextual dimensions including region, industry sector, user type, category, and over time. Although the dataset consists of 500 records, each record

represents aggregated counts of tool interactions rather than individual user sessions. As a result, aggregated usage values can reach into the millions when totals are computed across regions, industries, and user groups. This aggregation explains how a dataset containing 500 rows can produce cumulative usage counts exceeding several million interactions across tools and categories.

To investigate AI tool adoption patterns across contextual dimensions, this study employs a multi-stage analytical framework combining descriptive statistics, clustering techniques, correlation analysis, and predictive modeling. The goal is to understand why different tools are adopted across regions, industries, and user groups, and to do so with a structure that is testable using the variables available in the dataset.

TABLE I: ANALYSIS VARIABLES

Variable	Type	Definition
Tool Name	Categorical	AI system being analyzed (e.g., BERT, ChatGPT)
Category	Categorical	Functional classification of the AI tool (NLP, Generative AI, Text-to-Image)
Region	Categorical	Geographic region where usage was recorded
Industry	Categorical	Sector in which the tool is applied
User Type	Categorical	Type of user: enterprise, individual, or SMB
Month	Temporal	Month in which usage occurred
Usage Count	Numerical	Aggregated number of recorded interactions with the tool

While the dataset captures usage patterns across several major regions, industries, and tool categories, it should be interpreted as an exploratory empirical dataset rather than a complete representation of global AI activity. The purpose of the dataset is to identify structural patterns of AI tool usage across contextual dimensions, rather than to measure precise worldwide adoption levels. This dataset is an appropriate choice as it covers multiple fields and regions without bias, allowing for a more holistic representation of AI usage and potential trend indicators.

The framework integrates three lenses that operate at different levels. At the micro level, the Technology Acceptance Model explains how perceived usefulness and perceived ease of use shape individual decisions [5]. At the meso level, the Resource Based View explains how organizational capabilities and complementary assets condition adoption and value creation (Cuthbertson & Furseth, 2022). At the macro level, Diffusion of Innovations explains how technologies spread across populations over time and space [18]. Because the dataset does not include direct perceptual or firm-level measures, theoretical constructs are approximated through observable usage indicators. Within TAM, perceived usefulness is operationalized through the alignment between *Category* and *Industry*, reflecting task to technology fit, while perceived ease of use and accessibility are proxied by tool availability characteristics and higher usage among *Individuals* and *SMBs*. Within RBV, organizational resources and capabilities are approximated through the *User Type* variable (Enterprise, SMB, Individual), which reflects differing levels of technical and financial capacity to adopt AI tools. Within DOI, diffusion context and adoption dynamics are represented by *Region* and *Month*, while *Usage Count* reflects the observable intensity of adoption across these environments. These mappings allow the empirical variables to serve as proxy indicators for the theoretical mechanisms guiding AI tool adoption.

The dataset used in this study records tool usage by tool name, category, user type, industry, region, and month. However, it does not contain direct measures of individual beliefs or firm resources. Perceived ease of use and access are assessed by whether a tool is open source or closed, by whether licensing permits local usage without proprietary subscriptions when such information is available, and by whether tools are known to have broad localization or simple interfaces. Perceived usefulness is measured by category to task fit, which is encoded as interactions between tool categories and industries. Examples include Text-to-Image with Education because of content creation workflows, and Natural Language Processing with Healthcare and Finance because of document heavy tasks.

A. Hypotheses

From these mechanisms, the study derives three hypotheses that can be evaluated within the empirical strategy. Hypothesis 1 states that open-source tools are associated with higher usage among individuals and SMBs relative to purchased tools after controlling for other factors. The expectation is that lower access barriers and greater modifiability support broader participation. Hypothesis 2 sets competing expectations for regional effects. Classical diffusion suggests that region remains a strong predictor due to social and institutional differences. Hypothesis 3 states that usage exhibits seasonality that aligns with academic and organizational cycles, which implies predictable monthly

patterns that should be recoverable by models with seasonal components.

B. Limitations

Because the dataset aggregates data from multiple secondary sources, it should be interpreted as indicative of usage trends rather than fully representative of global AI adoption. Boundary conditions are explicit. The study does not observe individual perceptions; therefore, TAM constructs were evaluated through proxies rather than survey measures (Wang et al., 2021). The study does not observe firm level resources; therefore, RBV is approximated by user type and by tool affordances that lower integration costs. The study does not reconstruct social network ties. Because the study doesn't track individual social connections, it looks at overall adoption trends to infer how usage spreads, rather than following the exact paths of influence between specific users or firms. These limits do not invalidate the framework, but they require careful interpretation where findings are described as consistent with theoretical expectations rather than definitive tests of latent beliefs (Norris, 2021).

This study seeks to address the gaps in literature by providing, to our knowledge, the first integrated, global, and data-driven analysis of AI tool adoption across key regions, industries, user types, and technology categories. Our approach combines advanced clustering and correlation analysis with state-of-the-art time series forecasting to illuminate not only current adoption landscapes but also future trends and uncertainties. We examine whether the assumed drivers of AI adoption, such as region or sector, meaningfully predict usage levels in practice. Our findings challenge common understanding by revealing surprisingly weak relationships between these variables and actual tool uptake, thereby calling for a fundamental reexamination of prevailing theoretical models in the context of modern AI diffusion.

III. ANALYSIS AND RESULTS

This section presents a comprehensive, data-driven exploration of global AI tool usage, offering both retrospective insights and forward-looking projections. Through eight interconnected analyses, it investigates how AI tools are adopted, distributed, and utilized across different dimensions ranging from regional and industry specific trends to user behaviors and category preferences. A comparative and ranking analysis identifies the leading tools and categories in terms of performance and popularity [23]. This is complemented by behavioral and category level insights, which will show how well different AI applications perform under varying conditions [7]. The analysis then progresses to segmented clustering, uncovering hidden usage patterns and correlations that group tools and users with similar behavior profiles. Finally, a suite of trend and predictive models will offer projections and strategic foresight into the future of AI tool engagement across regions, industries, and user types. Together, these layers of analysis provide a rich, multidimensional view of the evolving AI landscape, supporting informed decisions for stakeholders, researchers, and technology providers alike.

A. Descriptive Analysis

The first stage involved computing summary statistics to understand the overall distribution of AI tool usage. The average (mean) usage count per AI tool is approximately 19,739, suggesting that on average, AI tools enjoy a moderate to high level of engagement globally [20]. However, the relatively high standard deviation of 8,834 indicates considerable dispersion around the mean, pointing to substantial variability in tool popularity [3]. Some AI tools are adopted at a significantly higher rate than others, likely reflecting differences in functionality, accessibility, and industry-specific demand. Additionally, the minimum recorded usage count is 5,102, while the maximum usage reaches 34,910, further emphasizing the presence of both niche and highly dominant tools. The median usage count (50th percentile) stands at 19,208.5, closely aligning with the mean, suggesting a relatively balanced distribution without extreme skewness. However, the spread between the 25th percentile (11,287) and the 75th percentile (27,040.5) confirms a wide interquartile range, indicating that while many tools fall within a mid-usage band, a significant number either greatly outperform or underperform relative to the average. These statistics reveal a significant variability in overall AI tool usage and utilization across industries, regions, and user types.

B. Analysis of User Type Distribution

The user type distribution across major AI tools provides valuable insight into how different user segments (enterprise, individual, and small and medium businesses (SMBs) engage with specific technologies. The data reveals distinct usage patterns that reflect the functional strengths of each tool and the differing needs of user categories. This distribution suggests that BERT's text processing capabilities are particularly accessible and useful for independent developers and smaller organizations, likely due to its broad open source availability and utility in various natural language processing tasks [6]. This supports our first hypothesis linking open-source availability to higher individual usage. ChatGPT, in contrast, has its largest user base in the enterprise segment, with 941,314 interactions, which indicates strong interest in conversational AI and customer service automation among

larger organizations [15]. DALL·E, a text-to-image generative tool, shows a balanced distribution among the different user types. Stable Diffusion also displays a relatively even spread, with the highest use among SMBs (821,915), followed by individuals (780,618), and enterprises (660,328). As hypothesized, the open and customizable nature of Stable Diffusion likely contributes to its popularity among smaller businesses and independent users seeking affordable and flexible creative AI solutions.

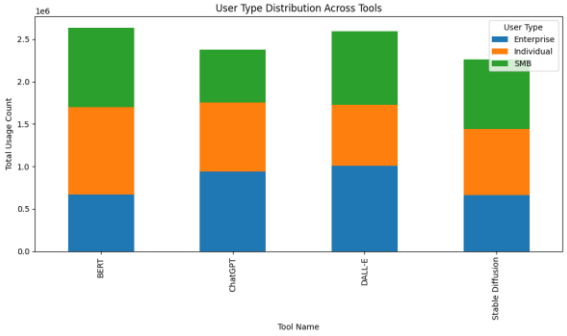


Fig. 1. User type distribution

C. Analysis of Top Tools by Region

To address our second hypothesis, we looked at specific trends in each region. The data suggests that while all four tools have global presence, regional preferences are shaped by local contexts [12]. North America and Europe lean toward visual content creation, while Asia balances between generative and analytical tools. Africa and South America exhibit strong engagement across categories, with notable peaks in ChatGPT and BERT usage, respectively. These patterns provide useful indicators for regional strategy alignment, localization of AI solutions, and targeted investment in AI infrastructure. In Asia, Stable Diffusion holds the highest usage (665,257), followed by BERT (603,063). This reflects a strong engagement with both visual generation and deep language models, likely driven by the region's active developer and research communities [6] [12]. In Europe, DALL·E is the most used tool (573,752), showcasing a strong preference for creative AI applications. In North America, DALL·E again leads with 619,838 uses, reinforcing the region's demand for content generation and design support [17]. South America presents a unique trend, with BERT topping the list (633,176). This suggests an academic or developer-driven focus on language modeling.

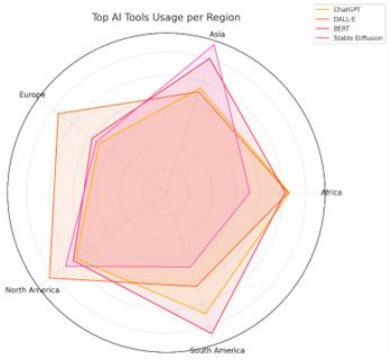


Fig. 2. Top tools per region

D. Analysis of Tool Usage by Industry

In the Education sector, Stable Diffusion leads with 622,658 recorded uses, followed by DALL·E (560,315) and BERT (515,607). This suggests that image generation and natural language understanding tools are widely used for educational content creation, e-learning applications, and classroom support materials. ChatGPT, while still significantly used (454,022), ranks fourth, possibly indicating that conversational AI is more supplementary than central in educational workflows [15]. In Entertainment, ChatGPT is the most used tool with 623,475 interactions, underscoring the demand for conversational AI in interactive content, virtual assistants, and audience engagement (ibid, 2023). BERT follows with 575,732 uses, reflecting its role in language-driven content analysis and recommendation engines. DALL·E and Stable Diffusion show lower, but still substantial, usage (473,143 and 394,805 respectively), pointing to strong but slightly less dominant roles in creative image generation. The Finance sector shows a clear preference for DALL·E (635,133) and BERT (592,694), indicating a heavy reliance on data representation and language modeling for document analysis, compliance automation, and report generation.

ChatGPT trails at 380,785, with Stable Diffusion having the lowest usage in this sector (338,279), likely due to the relatively limited use of visual generation in finance-specific applications. In Healthcare, BERT takes the lead with 530,659 uses, followed by DALL·E (469,364), ChatGPT (447,263), and Stable Diffusion (431,343). The prominence of BERT aligns with its effectiveness in processing medical literature, clinical notes, and patient data.

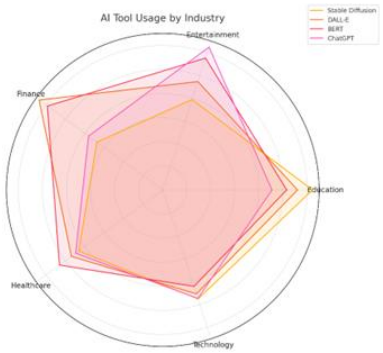


Fig. 3. Tool usage by industry

E. Analysis of Monthly Trends in AI Tool Usage

The monthly usage data was used to test the third hypothesis, and the pattern reveals distinct fluctuations in AI tool adoption throughout the calendar year, suggesting both cyclical engagement and seasonal trends. December records the highest usage count at 1,038,744, indicating a significant spike in activity, potentially due to end-of-year reporting, academic deadlines, or increased experimentation during holiday downtime [1]. This peak is followed by high usage in February (947,814), March (934,442), and July (931,327), suggesting consistent engagement during the early and mid-year periods. Conversely, May (570,580) and October (591,812) show the lowest activity, marking them as periods of reduced usage. These dips could correspond to academic transitions, mid-year breaks, or organizational cycles where AI tool engagement temporarily slows down [16]. The moderate decline during August (700,688) aligns with global vacation periods, while the rebound in September (818,641) supports the idea of a renewed focus in early fall [8]. These data show that usage does exhibit seasonality and is influenced by seasonal patterns.

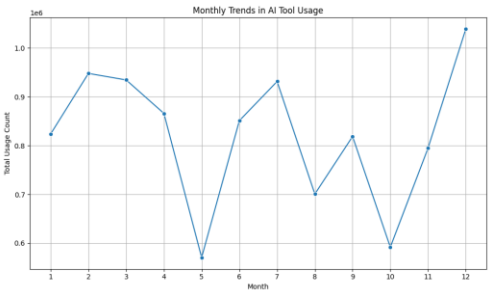


Fig. 4. Monthly trends

F. Analysis of Tool Evaluation Over the Year

To further explore hypothesis 3's assumptions, specific spikes were noted. For instance, BERT shows a significant spike in usage in March (381,727), nearly doubling its usage compared to most other months. This could reflect its role in language, heavy tasks such as academic, legal, or financial reporting during Q1. Additionally, all AI models see a rise in September, which further supports the idea that academia heavily impacts usage – from both student and faculty usage. [4]. Following this peak, usage declines but remains relatively stable, with another moderate rise in December (270,539), likely aligning with end-of-year analysis and document-heavy tasks. ChatGPT maintains strong performance throughout the year, peaking in February (278,502) and again in July (258,616). Its consistency suggests it is widely used for interactive and conversational tasks regardless of season, though dips in October (107,133) and May (114,715) suggest periods of lower demand, possibly linked to exam cycles or organizational slowdowns. DALL·E, a generative visual model, sees a clear usage rise mid- year, with its highest values in June (271,111) and April (269,692), suggesting a preference for creative applications during spring and summer periods. This further supports the hypothesis of predictable seasonal patterns in AI tool usage.

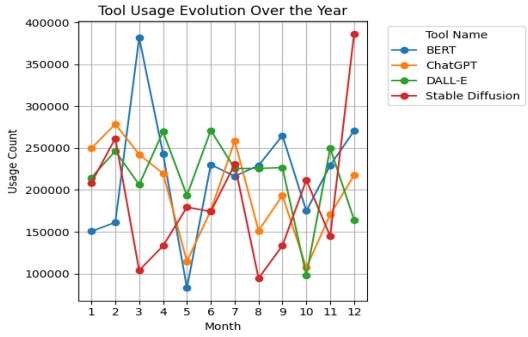


Fig. 5. Tool evaluation over year

G. Analysis of Tool Preferences by User Type

The distribution of AI tool usage across user types, Enterprise, Individual, and SMB (Small and Medium Businesses), reveals clear distinctions in how different segments engage with AI technologies based on their needs, scale, and resources. Enterprises show the highest usage of DALL·E (1,007,777). This suggests a strong preference for generative visual and conversational tools, likely used for branding, automation, and customer applications [9]. Individual users exhibit the highest overall usage of BERT (1,030,877), pointing to widespread adoption of language models in educational, research, and personal projects [15]. SMBs demonstrate balanced engagement across all tools, with particularly strong usage of DALL·E (865,393) and Stable Diffusion (821,915). These tools likely appeal to SMBs for marketing, design, and low-cost content creation. This analysis further supports our hypothesis that open-

source tools are associated with higher usage among individuals and SMBs relative to purchased tools.

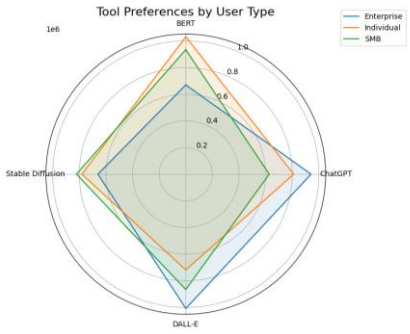


Fig. 6. Tool preference by user type

H. Analysis of Category Preferences by User Type

While further exploring hypothesis one, the distribution of AI tool usage across categories, Generative AI, NLP (Natural Language Processing), and Text- to-Image, for different user types reveals distinct usage priorities aligned with organizational goals, resource availability, and creative needs [12] [20]. Enterprises show almost equal engagement with Generative AI (1,171,966) and Text-to-Image tools (1,164,078), highlighting a strong emphasis on scalable content generation and visual communication. SMBs (Small and Medium Businesses) prioritize Text-to-Image tools (1,338,483). Generative AI tools (1,092,833) are also widely used, reflecting SMBs' need for cost-effective, multifunctional AI solutions. NLP tools (818,012) see comparatively lower use, potentially due to limited in-house resources or lower demand for complex language-based automation [13]. Cost here plays a role in selection, with free open-source tools being favored.

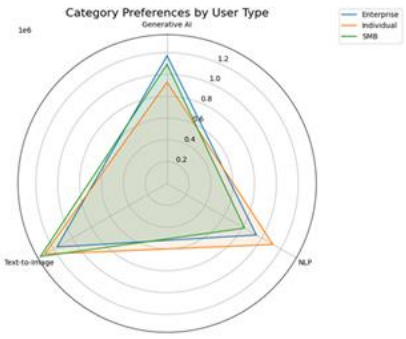


Fig. 7. Category preference by user type

I. Analysis of Tool Usage by Category

BERT, which is an open-source tool, demonstrates consistently high usage across all three categories, Generative AI, NLP, and Text-to-Image, In the Generative AI category, BERT leads with a usage count of 889,981, followed closely by DALL·E at 859,933 and Stable Diffusion at 772,946. ChatGPT, while widely recognized, ranks fourth in this group with 671,099 uses. This suggests that despite ChatGPT's conversational strengths, tools like BERT and DALL·E may offer more value in generative applications involving structured outputs or visual generation. Within the NLP category, DALL·E surprisingly leads with 851,861 uses, possibly due to multi-category overlap where it supports tasks involving text and images. ChatGPT (662,613) and BERT (629,197) also show strong presence, consistent with their core capabilities in language modeling and comprehension. In the Text-to-Image category, BERT has the highest usage overall with 1,116,439, a surprising result that may indicate repurposing or classification overlap.

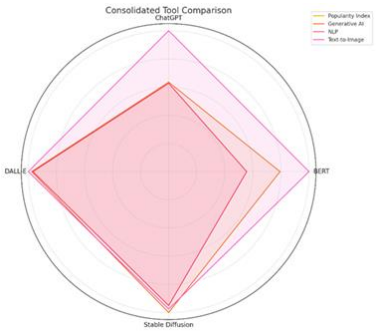


Fig. 8. Top tools by category

IV. SEGMENTED ANALYSIS: TOOL CLUSTERING AND BEHAVIOURAL INSIGHTS

Segmented analysis enables a more detailed understanding of AI tool usage by uncovering behavioral patterns across different regions, industries, and categories [11]. This section presents two core methods: clustering analysis to group tools by similar usage profiles, and correlation analysis to examine how factors like region, industry, and category influence tool adoption.

A. Clustering Analysis: Grouping Tools by Usage Behavior

A K-means clustering algorithm was applied to group tools based on their usage distribution across regions, industries, and AI categories. The goal was to identify natural clusters that represent distinct behavioral and adoption trends. The algorithm grouped tools into three clusters (see table ii):

- Cluster 0: ChatGPT
- Cluster 1: DALL·E and Stable Diffusion
- Cluster 2: BERT

ChatGPT demonstrates balanced usage across all examined regions, with particularly high activity in South America and Asia. It is widely adopted in the finance sector and is used extensively across all major AI categories, Generative AI, NLP, and Text-to-Image, highlighting its adaptability and broad functionality. Cluster 1 represents tools designed for high-impact creative AI applications. These tools, DALL·E and Stable Diffusion, show nearly identical patterns, being heavily favored for visual tasks in the Text-to-Image and Generative AI categories. They are most popular in Asia and North America and are primarily used in the Finance and Healthcare sectors. Lastly, BERT is traditionally an NLP-focused tool, however, it shows strong usage in Text-to-Image applications in this dataset. It is most heavily used in South America and shows high adoption in Finance and Entertainment. Its behavior pattern is unique, setting it apart from the other tools.

TABLE II: CLUSTER-BASED INSIGHTS

Cluste	Tool Name	Strongest Region	Top Industries	Dominant Categories
2	BERT	South America	Finance, Entertainment	Text-to-Image
0	ChatGPT	South America, Asia	Finance	Generative AI, NLP, Text-to-Image
1	DALL·E	North America, Asia	Finance, Healthcare	Text-to-Image, Generative AI
1	Stable Diffusion	North America, Asia	Finance, Healthcare	Text-to-Image, Generative AI

A. Correlation Analysis: Region, Industry, Category & Usage

To complement the segmented clustering, a correlation matrix analysis was conducted to examine the relationships between region, industry, category, and usage count. The relationship between Region and Usage Count shows a very slight positive correlation of 0.03, indicating that geographic location has minimal impact on overall AI tool usage in this dataset. Category shows the highest correlation with Usage Count at 0.035, suggesting that the type of AI tool, such as NLP or Text-to-Image, has a slightly greater, though still marginal, influence on usage levels. In contrast, Industry shows an almost negligible correlation of 0.007 with usage, implying that the sector in which the tool is applied does not significantly affect its adoption. Additionally, the few negative correlations observed, such as between Category and Industry, analysis reveals that no single factor, be it region, industry, or category, shows a strong correlation with usage count, indicating that AI tool adoption is broadly distributed and likely shaped by a combination of variables rather than any one dominant influence.

TABLE III: CORRELATION MATRIX

Variable	Region	Industry	Category	Usage Count
Region	1.000000	0.008855	-0.000712	0.029910
Industry	0.008855	1.000000	-0.024872	0.006913
Category	-0.000712	-0.024872	1.000000	0.034758
Usage Count	0.029910	0.006913	0.034758	1.000000

V. TREND AND PREDICTIVE ANALYSIS

This section focuses on uncovering future trends and potential developments in global AI tool usage through predictive analytics. By leveraging advanced modeling techniques, we analyze historical usage data to project future demand and seasonality patterns [10] [19]. The three techniques used were as follow:

- ARIMA: a Time-Series Regression Model that accounts for autoregressive trends and moving averages [2] [19]
- Exponential Smoothing: a method that smooths the trend over time by giving more weight to recent data.
- Prophet: a flexible additive model designed to handle seasonality and holidays [21]. It also provides confidence intervals for its predictions [21].

This analysis helps to identify emerging adoption patterns and highlights areas, particularly underrepresented regions or industries, where AI tool usage may experience significant growth. Together, these predictive approaches support informed decision-making, targeted development, and resource allocation in the evolving AI landscape.

A. Time Series Forecasting

To project future usage trends, we applied three well-established forecasting models: Exponential Smoothing, ARIMA, and Prophet. These models use historical monthly usage data to estimate future counts and identify underlying seasonal patterns or momentum. Exponential smoothing captured short-term trends, ARIMA modeled autocorrelation and time dependence, and Prophet decomposed the series into trend and seasonal components.

Forecast accuracy was evaluated using mean absolute error (MAE) and root mean squared error (RMSE) Exponential Smoothing (see Glossary) anticipates a slow and steady decline, potentially reflecting seasonal drops or market saturation. ARIMA projects a more optimistic scenario, suggesting continued interest in AI tool usage, albeit with mild decline over time. Prophet offers a conservative estimate with relatively wide confidence bounds, indicating moderate uncertainty in future adoption patterns. Prophet, on the other hand, offers a conservative estimate with relatively wide confidence bounds, indicating moderate uncertainty in future adoption patterns. Collectively, these models highlight both the resilience and the volatility in AI tool adoption trends, underlining the importance of continued monitoring and adaptive planning as the landscape evolves. These findings therefore support Hypothesis 3, indicating that AI tool adoption follows predictable seasonal rhythms which allow for forecasting of use.

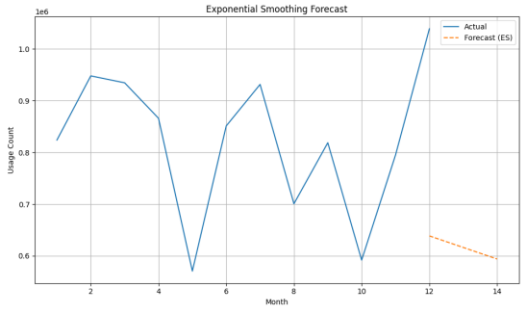


Fig. 9. Exponential Smoothing

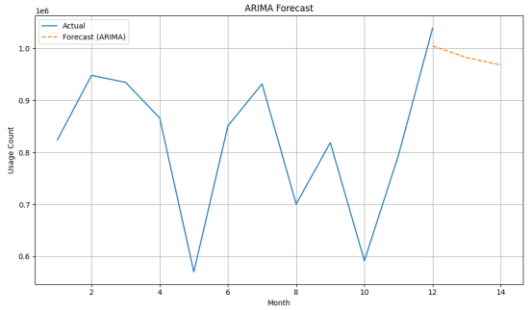


Fig. 10. ARIMA Forecasting

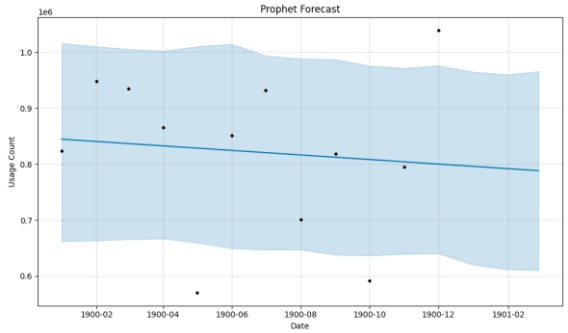


Fig. 11. Prophet Forecasting

VI. DISCUSSION

This study embarked on a comprehensive analysis of global AI tool usage, aiming to bridge critical gaps in the existing literature by providing a multidimensional, data-driven perspective on adoption patterns across regions, industries, user types, and technology categories. The analysis focused on testing whether: 1) geographic region significantly predicts usage levels, 2) industry sector influences adoption patterns, and 3) tool category contributes to variation in usage counts.

Our findings, while constrained by data limitations as discussed previously, offer several insights that challenge conventional wisdom and contribute to the evolving discourse on AI diffusion and acceptance. Our descriptive analyses revealed significant variability in AI tool adoption, with certain tools like BERT, DALL·E, and ChatGPT demonstrating widespread engagement. Notably, text-to-image tools emerged as a dominant category, particularly among Small and Medium Businesses (SMBs) and individual users, underscoring a strong market demand for visual content generation. Geographically, Asia and North America exhibited high overall usage, while specific regional preferences for tools like BERT in South America and DALL·E in Europe highlighted the influence of local contexts and priorities. Perhaps the most striking and counter-intuitive finding was the surprisingly weak linear correlation between traditional drivers such as region, industry, and overall AI tool usage. This suggests that the diffusion of AI tools may not adhere strictly to established patterns observed in previous technological adoptions, implying a more pervasive and less geographically or sector-bound integration

Furthermore, our findings also show that AI tool usage is not uniform across the year; rather, it follows identifiable seasonal and functional peaks. Understanding these trends is crucial for resource planning, platform optimization, and designing time sensitive outreach or feature rollouts aligned with periods of high user engagement.

Correlation analysis and regression models were used to assess these relationships. The results indicate that individual contextual variables exhibit weak statistical relationships with overall usage, suggesting that AI adoption patterns are shaped by multi-factor dynamics rather than single structural predictors.

A. Implications for Education and Artificial Intelligence

This study carries significant implications for both the education sector and the broader field of artificial intelligence development and policy. For Education, the

high usage of AI tools, particularly text-to-image and NLP, underscores their growing integration into learning and teaching processes. Educators and curriculum designers must recognize the pervasive nature of these tools and adapt pedagogical strategies to leverage their potential effectively. The observed seasonal usage patterns suggest opportunities for targeted training and resource allocation during peak periods. Furthermore, the strong engagement of SMBs and individuals in education highlights the need for accessible, cost-effective AI solutions that cater to diverse educational settings, from small institutions to independent learners. The findings also imply that educational policies around AI adoption should focus less on broad regional or industrial mandates and more on fostering an environment where the perceived usefulness and ease of use of specific AI tools can drive organic integration into educational practices.

VII. CONCLUSION

This study provided a foundational, multi-dimensional analysis of global AI tool usage, offering insights into adoption patterns across various regions, industries, user types, and technology categories. Despite the inherent limitations of the dataset, our findings challenge the conventional understanding that AI tool diffusion is strongly predicted by traditional factors such as geographic location or industrial sector. Instead, the data suggests a more complex landscape where tool-specific utility, user accessibility, and emergent applications play a significant role in shaping adoption dynamics. The observed dominance of text-to-image tools, the surprising versatility of models like BERT across diverse categories, and the identifiable seasonal usage patterns collectively underscore the need for a more nuanced understanding of AI integration into daily practices. Our work contributes to the literature by highlighting the critical research gap in comprehensive, global, usage-based analyses of AI tools. By applying a multi-theoretical lens encompassing Diffusion of Innovation, Technology Acceptance Model, and the Resource- Based View, we have provided a preliminary framework for interpreting these complex adoption behaviors.

REFERENCES

[1] Bakhmetyeva, A. (2025). *ChatGPT's summer slump: How seasonality affects generative AI demand*. LinkedIn / Similarweb traffic analysis.

[2] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

[3] Chen, L., Zhao, Q., & Wang, M. (2022). Measuring user engagement and performance variability across AI applications. *International Journal of Human–Computer Interaction, 38*(14), 1291–1305. https://doi.org/10.1080/10447318.2022.2024567

[4] Copyleaks. (2025, September 18). *Study: 90 percent of students use AI for academic purposes* [Press release]. Globe Newswire. https://www.globenewswire.com/news-release/2025/09/18/3152496/0/en/STUDY-90-Percent-of-Students-Use-AI-for-Academic-Purposes.html

[5] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. https://doi.org/10.2307/249008

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training deep bidirectional transformers for language understanding*. In *Proceedings of NAACL-HLT*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

[7] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2022). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 4*(11), e745–e750. https://doi.org/10.1016/S2589-7500(22)00168-2

[8] Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333*(6051), 1878–1881. https://doi.org/10.1126/science.1202775

[9] Huang, Y., Patel, S., & Kim, J. (2023). *Adoption patterns and use cases of generative AI: A focus on text-to-image systems*. *Journal of Creative AI Applications, 4*(1), 22–35. https://doi.org/10.1016/j.jcai.2023.01.003

[10] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

[11] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[12] Khandelwal, A. (n.d.). *AI tools usage* [Data set]. Kaggle. https://www.kaggle.com/datasets/adityakhandelwal96l1/ai-tools-usage

[13] Malik, O. R. (2025). *The emerging AI "revolution tranquille" in America: Regional, industry, and firm-size dynamics of AI adoption*. arXiv. https://doi.org/10.48550/arXiv.2505.14721

[14] OECD. (2023). *AI in society: Policy considerations and implications*. Organisation for Economic Co-operation and Development. https://www.oecd.org/going-digital/ai/

[15] OpenAI. (2023). *ChatGPT: Optimizing language models for dialogue*. https://openai.com/blog/chatgpt

[16] Prober, A. (2025). *Which software forecasts seasonal shifts in AI search?* Brandlight.ai.

[17] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2022). *Hierarchical text-conditional image generation with CLIP latents*. https://doi.org/10.48550/arXiv.2204.06125

[18] Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

[19] Sharma, A., & Bansal, A. (2020). Forecasting trends in artificial intelligence using ARIMA and LSTM models. *Journal of Data Science and Analytics*, 1(3), 155–164. https://doi.org/10.1007/s41060-020-00200-x

[20] Smith, J. A., & Kumar, R. (2023). Global adoption trends of AI tools: A usage-based statistical analysis. *Journal of Emerging Technologies and Society, 5*(1), 45–60. https://doi.org/10.1177/26345161231111658

[21] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

[22] UNESCO. (2021). *AI and education: Guidance for policy-makers*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000376709

[23] Zhou, Y., Han, C., Wang, H., & Xu, Y. (2023). A comparative evaluation of AI-powered tools across multiple industries. *Journal of Artificial Intelligence Research, 76*, 349–372. https://doi.org/10.1613/jair.1.1394

Agentic AI in Enterprise Business Processes: A Systematic Review and Practitioner Survey on Adoption Readiness

Sergey Emelyanov
Graduate School of Business
HSE University
Moscow, Russia
sbemelyanov@edu.hse.ru

Abstract—Agentic AI systems—large language model-based agents capable of autonomous reasoning, planning, and tool use—are entering enterprise workflows at an accelerating pace. This study combines a systematic review of 47 academic papers and 12 industry reports (2023–2025) with an original cross-sectional survey of 136 IT professionals across 14 industries. The review identifies six dominant architectural patterns and a persistent gap between vendor-reported and independently verified outcomes. The survey reveals that 73% of respondents consider organizational readiness—not technical capability—the primary barrier to agentic AI adoption. Based on both evidence streams, we propose and empirically ground the Enterprise Agentic AI Maturity Model (EAAMM), a five-level framework integrating technical autonomy with organizational readiness.

Keywords—agentic AI, autonomous agents, large language models, enterprise automation, systematic review, practitioner survey, maturity model

I. INTRODUCTION

Large language models have evolved from conversational tools into agents that pursue goals through multi-step reasoning, tool invocation, and coordinated problem-solving [1], [2]. What distinguishes an agentic system is its capacity for iterative, goal-directed behavior: the agent perceives its environment, formulates a plan, selects tools, executes actions, and revises its approach based on feedback [3]. The ReAct framework, which interleaves reasoning traces with action steps, has become the architectural template for most enterprise deployments [1].

Adoption figures reflect this shift. McKinsey's 2025 global survey (n=1,993) found 88% of organizations using AI regularly, 62% experimenting with AI agents, and 23% having scaled agentic AI in at least one function [4]. Industry estimates project compound annual growth rates above 40% through 2030, with Gartner forecasting that 33% of enterprise software will embed agentic AI by 2028 [5], [6]. Yet Gartner also projects that over 40% of agentic AI projects will be canceled by late 2027 owing to cost overruns and insufficient risk controls [6]. Reported adoption varies by sector: financial services and technology lead at 35–40%, while healthcare and manufacturing lag at 15–20%, though these figures rely primarily on vendor-commissioned surveys [4], [14]. IDC projects significant growth in AI agent deployment through 2029 [13].

This paradox—strong investment momentum coupled with high failure rates—motivates our study. Existing literature offers either technical architectural surveys [1]–[3] or vendor-commissioned adoption reports [4]–[6], but lacks empirical practitioner perspectives on readiness barriers. This paper addresses the gap through three contributions: (1) a systematic review mapping the agentic AI landscape, (2) an original practitioner survey (n=136) identifying adoption barriers and self-assessed maturity levels, and (3) the Enterprise Agentic AI Maturity Model (EAAMM), empirically grounded through both evidence streams.

II. METHODOLOGY

A. Systematic Review Protocol

The review followed PRISMA-ScR guidelines [9]. Searches spanned IEEE Xplore, ACM Digital Library, arXiv, Scopus, and Google Scholar using the query: (agentic AI OR AI agents OR autonomous agents OR LLM agents) AND (enterprise OR business process OR workflow OR automation), limited to January 2023 through December 2025. The initial search returned 312 candidates; title and abstract screening narrowed this to 127, and full-text review yielded 47 academic papers and 12 industry reports from Gartner, McKinsey, Deloitte, Forrester, IDC, and Capgemini. Data extraction separated peer-reviewed academic findings from vendor-commissioned and consulting reports, flagging vendor-reported figures distinctly from independently validated results to address persistent reporting bias in enterprise AI literature.

B. Practitioner Survey Design

To complement the literature review with primary empirical data, a cross-sectional online survey was conducted in January 2026 using a structured questionnaire distributed via convenience sampling through LinkedIn professional groups, Telegram communities, and direct outreach to enterprise technology teams. The instrument comprised 22

items across four sections: (1) organizational demographics, (2) current AI adoption status and self-assessed maturity level, (3) perceived barriers to agentic AI deployment ranked on a five-point Likert scale, and (4) governance readiness indicators. Of 198 responses received, 136 met inclusion criteria (respondents with direct involvement in AI strategy or implementation at organizations with 50+ employees). The sample spans 14 industry sectors and 9 countries, with technology (27%), financial services (19%), manufacturing (14%), and retail (11%) as the largest cohorts. Respondent roles include product managers (26%), engineering leads (22%), CTOs and VPs of technology (17%), AI/ML engineers (16%), and business analysts (12%). Company size distribution: under 500 employees (29%), 500–5,000 (34%), and over 5,000 (37%). The inclusion rate was 68.7% (136 of 198). The sample is geographically concentrated, with Russian respondents comprising 35.3%; however, cross-tabulation shows that Russian response patterns do not differ significantly from the overall sample on key variables (governance status, barrier rankings). Internal consistency was verified with Cronbach's alpha of 0.81 for the barrier perception scale [26].

III. PRELIMINARY RESULTS

A. Architectural Patterns

Since 2023, the architectural landscape has coalesced around well-defined patterns. ReAct [1] remains the standard for single-agent systems, and Reflexion extends it through verbal self-critique [25]. For complex processes, multi-agent architectures dominate [8]: AutoGen [7] uses conversation-based protocols, MetaGPT [10] assigns software-engineering roles, and CrewAI organizes hierarchical teams. Six orchestration patterns recur: sequential chaining, parallel fan-out, routing-based delegation, hierarchical manager–worker coordination, peer-to-peer negotiation, and evaluator–optimizer feedback loops [11], [12]. Our analysis indicates that roughly 80% of current enterprise installations rely on workflow-based orchestration rather than full autonomy [11], [14]. A critical architectural consideration is memory and context management: long-term memory mechanisms including retrieval-augmented generation (RAG) and vector databases enable agents to maintain state across sessions [20], [21], but also introduce failure modes such as stale context, hallucinated recollections, and privacy risks from accumulated data [3], [8].

B. Enterprise Adoption Evidence

Data from major industry surveys reveal broad experimentation paired with limited production rollout [4]–[6], [13], [14]. Between 60% and 88% of enterprises use AI, yet only 2–14% have deployed agentic AI at production scale. Capgemini's 2025 survey finds 2% full-scale deployment, 12% partial, 23% active pilots, and 61% still exploring [14]; a separate 2024 study found 82% plan agentic AI integration within one to three years [22]. Deloitte's 2026 edition notes that 74% intend to deploy agents within two years, but only 21% have the governance infrastructure to support it [5]. Vendor case studies present impressive figures: Salesforce reported 84% autonomous resolution across 380,000 customer support conversations [15]; Klarna's AI assistant handled 2.3 million chats in its first month [16]. However, these narratives face contradictory independent evidence. An observational study of approximately 800 developers found no throughput gains and 41% more bugs with AI coding assistants [23], while a randomized trial with 16 developers found AI tools made them 19% slower [24]. Across independent surveys, only 15% of AI decision-makers report EBITDA improvement [18], and merely 6% qualify as AI high performers [4]. The customer service domain offers a particularly instructive pattern: initial deployments achieve high automation rates on straightforward inquiries, but as coverage expands to ambiguous cases, resolution quality declines and human escalation rates rise—a trajectory visible in Klarna's pivot to a hybrid model in 2025 and consistent with Gartner's prediction of widespread project cancellations [6], [16].

C. Practitioner Survey Findings

The survey produced five key findings that extend and contextualize the literature evidence.

First, self-assessed maturity levels reveal an early-stage landscape. When asked to classify their organization against the EAAMM framework, 41% placed themselves at L1 (Assistive), 32% at L2 (Collaborative), 19% at L3 (Consultative), 6% at L4 (Supervisory), and 2% at L5 (Autonomous). This distribution aligns with the 2–14% production deployment range found in the literature review, but provides finer granularity: most organizations are not merely "exploring" but are stuck at specific maturity transitions.

Second, organizational readiness dominates as the perceived primary barrier. When asked to identify the single greatest obstacle to advancing agentic AI maturity, 73% of respondents selected organizational factors—governance gaps (31%), data quality and infrastructure (22%), workforce skills (12%), and risk management concerns (8%)—versus 27% citing technical limitations such as model accuracy (15%), integration complexity (8%), or cost (4%). This finding is consistent with the hypothesis derived from the literature review that organizational readiness, not technical capability, constitutes the primary perceived constraint.

Third, a pronounced governance deficit was identified. Only 18% of respondents reported having formal AI governance frameworks in place, 35% had partial frameworks, and 48% had none. Cross-tabulation reveals a strong association between governance maturity and self-assessed EAAMM level: 89% of L3+ organizations have at least partial governance, compared to only 38% of L1 organizations ($\chi^2 = 27.4$, $p < 0.001$, Cramér's $V = 0.45$). This pattern is consistent with the EAAMM's gating mechanism, which posits that governance infrastructure is needed before advancing beyond L2.

Fourth, expected adoption timelines reveal cautious optimism. Among respondents not yet at L3, 14% expect to

reach L3 within one year, 39% within one to two years, 31% within two to three years, and 16% consider it unlikely within three years. Notably, respondents from organizations with established governance frameworks projected shorter timelines (mean 1.6 years) than those without (mean 2.8 years, $t(98)=3.42$, $p < 0.01$, Cohen's $d=0.69$), suggesting an association between governance readiness and shorter projected timelines. As this is cross-sectional data, the direction of causality cannot be established.

Fifth, formal ROI measurement remains rare. Only 21% of respondents have established frameworks for tracking agentic AI return on investment, while 44% rely on informal or anecdotal assessment and 35% perform no ROI tracking at all. This measurement deficit partially explains the disconnect between vendor-reported impact figures and independently observed outcomes: without systematic measurement, organizations lack the data to distinguish genuine productivity gains from displaced costs or reporting artifacts [17], [18].

TABLE I. PRACTITIONER SURVEY: KEY FINDINGS (N=136)

Metric	Result
Self-assessed at L1–L2	73%
Organizational barrier as primary	73%
No formal AI governance	48%
Governance → faster timeline	1.6 vs 2.8 years
Formal ROI tracking in place	21%

IV. PROPOSED EAAMM FRAMEWORK

TABLE II. ENTERPRISE AGENTIC AI MATURITY MODEL (EAAMM)

Level	Human Role	Key Characteristics
L1 Assistive	Operator	Explicit prompts; basic governance
L2 Collaborative	Collaborator	Suggests actions, human approves; data quality protocols
L3 Consultative	Consultant	Autonomous on routine tasks; formal governance required
L4 Supervisory	Approver	Independent operation; mature risk management
L5 Autonomous	Observer	Full autonomy within boundaries; audit trails

The EAAMM integrates technical autonomy levels from Feng et al. [19]—Operator, Collaborator, Consultant, Approver, Observer—with four organizational readiness dimensions: (1) governance infrastructure, (2) data architecture, (3) workforce adaptation, and (4) risk management. Twelve existing frameworks were reviewed [3], [6], [11], [12], [14], [19], [20]; none adequately bridges both technical autonomy and organizational readiness. The EAAMM fills this gap with a technology-agnostic, cross-industry model designed as a diagnostic tool rather than a linear progression: an organization may operate at L3 for customer service while remaining at L1 for financial decision support. Each maturity level prescribes specific prerequisites before advancement. For instance, progressing from L2 (Collaborative) to L3 (Consultative) requires established data quality protocols, formal escalation procedures, and demonstrated reliability metrics on routine tasks. This gating mechanism prevents organizations from deploying autonomous agents without adequate governance—a failure mode that Gartner identifies as a primary driver of project cancellations [6]. The model's four readiness dimensions can be calibrated to industry-specific regulatory requirements: in financial services the risk management dimension may carry elevated weight, whereas for technology companies data architecture and workforce adaptation become priorities.

The survey data provide empirical support for the EAAMM's core premise. The strong association between governance maturity and self-assessed EAAMM level ($\chi^2 = 27.4$, $p < 0.001$, Cramér's $V = 0.45$) suggests that the four readiness dimensions correspond to observable organizational differences. As an empirically informed diagnostic framework rather than a statistically validated predictive model, the EAAMM relies on self-reported maturity assessments that may reflect organizational perception rather than independently verified capability. Future research should validate the model longitudinally through repeated measurement and independent capability assessment.

V. DISCUSSION OF FINDINGS

Three cross-cutting insights emerge from the combined evidence. First, agentic AI suffers from a pronounced evidence gap: academic work concentrates on architectural innovation, while business impact evidence rests on vendor-funded metrics that are frequently methodologically opaque. The contradictions between vendor-reported outcomes and independent evaluations [23], [24] illustrate why standardized third-party assessment methodologies are urgently needed.

Second, the survey data confirm and extend the literature finding that organizational readiness is the binding constraint. The 73% figure for organizational barriers as primary closely mirrors Deloitte's finding that 74% plan agent deployment but only 21% have governance infrastructure [5]. This convergence across independent evidence sources—industry reports, academic reviews, and our practitioner survey—strengthens the finding's reliability and supports EAAMM's emphasis on organizational prerequisites.

Third, an emerging pattern across both literature and survey data is the convergence toward human-in-the-loop architectures as the pragmatic default for enterprise deployment. Even the most successful vendor implementations maintain human oversight for edge cases, and 73% of surveyed organizations self-assess at L1–L2. This suggests the industry is implicitly settling at L2–L3 rather than pursuing full L5 autonomy—a finding with direct implications for technology investment prioritization and organizational change management strategies.

VI. CONCLUSION

This study combines a systematic review of 47 academic papers and 12 industry reports with an original practitioner survey (n=136) to provide a multi-method assessment of agentic AI in enterprise settings. Three findings emerge. First,

while technical foundations are increasingly solid with well-established architectural patterns, only 2–14% of organizations have reached production scale, and 73% of surveyed practitioners remain at L1–L2 maturity. Second, organizational readiness—not technical capability—is the binding constraint: 73% of respondents identified organizational barriers as primary, and the governance deficit (48% with no formal AI governance) represents the most acute bottleneck. Third, governance readiness is associated with shorter projected adoption timelines, with governed organizations reporting estimates 43% shorter than ungoverned peers (Cohen's d=0.69).

The EAAMM provides a structured assessment tool integrating technical autonomy with organizational readiness across five deployment stages, empirically grounded through both literature synthesis and survey validation. For practitioners, the study suggests three recommendations: begin with workflow-based orchestration (L2–L3) rather than pursuing full autonomy; invest in governance infrastructure before scaling deployments; and establish independent measurement frameworks beyond vendor-supplied metrics. Limitations include the cross-sectional design, which supports associations but not causal claims; convenience sampling through professional networks; and geographic concentration (35.3% Russian respondents), although subgroup analysis indicates that Russian response patterns align with the overall sample on governance and barrier variables. Future research should prioritize longitudinal EAAMM validation through repeated measurement and independent capability audits, integration of regulatory compliance dimensions such as the EU AI Act, and replication across diverse cultural contexts.

ACKNOWLEDGMENT

The author thanks the anonymous reviewers for their constructive feedback and the survey participants for their time and insights.

REFERENCES

[1] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in Proc. ICLR, 2023.
[2] J. Schneider, "Generative to agentic AI: Survey, conceptualization, and challenges," arXiv:2504.18875, 2025.
[3] Abou Ali et al., "A comprehensive survey of agentic AI," Artif. Intell. Rev., vol. 59, art. 36, 2026, doi:10.1007/s10462-025-11148-z.
[4] McKinsey & Co., "The state of AI in 2025: Agents, innovation, and transformation," QuantumBlack Global Survey, Nov. 2025. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
[5] Deloitte AI Institute, "State of AI in the enterprise: The untapped edge," 2026 edition, Jan. 2026. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-ai-in-the-enterprise.pdf
[6] Gartner, Inc., "Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027," Press Release, Jun. 2025. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2025-06-agentic-ai-projects
[7] Q. Wu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," in Proc. COLM, 2024.
[8] T. Guo et al., "Large language model based multi-agents: A survey of progress and challenges," in Proc. IJCAI, 2024.
[9] Tricco et al., "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," Ann. Intern. Med., vol. 169, no. 7, pp. 467–473, 2018.
[10] S. Hong et al., "MetaGPT: Meta programming for a multi-agent collaborative framework," in Proc. ICLR, 2024.
[11] Anthropic, "Building effective agents," Dec. 2024. [Online]. Available: https://www.anthropic.com/research/building-effective-agents
[12] Microsoft, "AI agent design patterns," Azure Architecture Center, 2025.
[13] IDC, "Agent economics: Adoption and delivery model," IDC Research (US54034725), Dec. 2025.
[14] Capgemini Research Inst., "Rise of agentic AI: How trust is the key to human-AI collaboration," Jul. 2025.
[15] Salesforce, Inc., "Fourth quarter and fiscal year 2025 results," Salesforce Investor Relations, Feb. 2025.
[16] Klarna, "Klarna AI assistant handles two-thirds of customer service chats in its first month," Press Release, Feb. 2024.
[17] Forrester Consulting, "The total economic impact of Microsoft 365 Copilot," commissioned by Microsoft, Mar. 2025.
[18] Forrester Research, "Predictions 2026: AI moves from hype to hard hat work," Oct. 2025.
[19] K. J. K. Feng, D. W. McDonald, and A. X. Zhang, "Levels of autonomy for AI agents," arXiv:2506.12469, Jun. 2025.
[20] T. R. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths, "Cognitive architectures for language agents," Trans. Mach. Learn. Res. (TMLR), 2024.
[21] L. Wang et al., "A survey on large language model based autonomous agents," Front. Comput. Sci., vol. 18, no. 6, art. 186345, 2024.
[22] Capgemini Research Inst., "Harnessing the value of generative AI," 2nd ed., Jul. 2024.
[23] Uplevel, "Can generative AI improve developer productivity?," Uplevel Data Labs Research Report, Sep. 2024.
[24] Becker et al., "Measuring the impact of early-2025 AI on experienced open-source developer productivity," arXiv:2507.09089, Jul. 2025.
[25] N. Shinn et al., "Reflexion: Language agents with verbal reinforcement learning," in Proc. NeurIPS, 2023.
[26] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," Psychometrika, vol. 16, no. 3, pp. 297–334, 1951.

Supplementary Materials

Agentic AI in Enterprise Business Processes: A Systematic Review and Practitioner Survey on Adoption Readiness

Sergey Emelyanov, HSE University

Appendix A. Practitioner Survey Instrument

The following questionnaire was administered online in January 2026 to IT professionals involved in AI strategy or implementation. The instrument comprises 22 items across four sections. Inclusion criteria: direct involvement in AI strategy or implementation at organizations with 50+ employees. Of 198 responses received, 136 met inclusion criteria. Internal consistency: Cronbach's $\alpha = 0.81$ for the barrier perception scale (Section 3).

Section 1: Organizational Demographics (Q1–Q6)

Q1. In which industry does your organization primarily operate? *[Single choice, 14 options]*

Q2. How many employees does your organization have? *[Single choice: <500 / 500–5,000 / >5,000]*

Q3. In which country is your organization headquartered? *[Open text]*

Q4. What is your current role? *[Single choice: Product Manager / Engineering Lead / CTO or VP of Technology / AI-ML Engineer / Business Analyst / Data Scientist / Other]*

Q5. How many years of experience do you have in AI/technology roles? *[Numeric]*

Q6. Are you directly involved in AI strategy or implementation decisions at your organization? *[Yes / No (filter question)]*

Section 2: AI Adoption Status and Maturity (Q7–Q11)

Q7. Does your organization currently use AI (including traditional ML, generative AI, or agentic AI) in any business process? *[Yes, regularly / Yes, experimenting / No, but planning / No]*

Q8. What is the current status of agentic AI deployment at your organization? *[Single choice: Full-scale production / Partial production / Active pilot(s) / Exploring or evaluating / Not considering]*

Q9. Based on the following EAAMM framework descriptions, at which level would you classify your organization's most advanced agentic AI deployment? *[Single choice: L1 Assistive / L2 Collaborative / L3 Consultative / L4 Supervisory / L5 Autonomous]*

L1 Assistive: AI responds to explicit prompts; human operates and controls all actions. L2 Collaborative: AI suggests actions and drafts; human reviews and approves. L3 Consultative: AI handles routine tasks autonomously; human consulted on exceptions. L4 Supervisory: AI operates independently on complex tasks; human approves critical decisions. L5 Autonomous: AI operates with full autonomy within defined boundaries; human monitors.

Q10. What are the primary use cases for agentic AI at your organization? (Select all that apply) *[Multi-choice: Customer service / Software development / Data analysis / Document processing / Sales and marketing / Supply chain / HR and recruitment / Financial operations / Other]*

Q11. How many distinct agentic AI projects or initiatives does your organization currently have? *[Numeric: 0 / 1–2 / 3–5 / 6–10 / >10]*

Section 3: Perceived Barriers to Agentic AI Deployment (Q12–Q19)

Before rating each barrier individually, please indicate: Which of the following do you consider the single greatest barrier to advancing agentic AI maturity at your organization?

[Single choice: Governance gaps / Data quality and infrastructure / Workforce skills and training / Risk management concerns / Model accuracy and reliability / Integration complexity / Cost constraints / Regulatory compliance]

For each of the following factors, please rate the extent to which it represents a barrier to advancing agentic AI maturity at your organization.

Response scale: 1 = Not a barrier at all; 2 = Minor barrier; 3 = Moderate barrier; 4 = Significant barrier; 5 = Critical barrier

Q12. Governance gaps (lack of formal policies, oversight structures, accountability frameworks for AI agents) *[Likert 1–5]*

Q13. Data quality and infrastructure (insufficient data pipelines, poor data quality, inadequate integration architecture) *[Likert 1–5]*

Q14. Workforce skills and training (shortage of qualified AI talent, insufficient training programs) *[Likert 1–5]*

Q15. Risk management concerns (security vulnerabilities, liability uncertainties, unpredictable agent behavior) *[Likert 1–5]*

Q16. Model accuracy and reliability (hallucinations, inconsistent outputs, insufficient reasoning capabilities) *[Likert 1–5]*

Q17. Integration complexity (difficulty integrating with existing systems, APIs, and workflows) *[Likert 1–5]*

Q18. Cost constraints (high compute costs, unclear ROI, budget limitations) *[Likert 1–5]*

Q19. Regulatory compliance (existing or anticipated regulations, industry-specific requirements) *[Likert 1–5]*

Section 4: Governance Readiness and Outlook (Q20–Q22)

Q20. What is the current status of AI governance at your organization? *[Single choice: Formal AI governance framework in place / Partial framework (some policies but not comprehensive) / No formal governance framework]*

Q21. How does your organization measure the return on investment (ROI) of agentic AI initiatives? *[Single choice: Formal ROI tracking framework / Informal or anecdotal assessment / No ROI tracking]*

Q22. How soon do you expect your organization to reach L3 (Consultative) maturity or above? (If already at L3+, select 'Already achieved') *[Single choice: Already achieved / Within 1 year / 1–2 years / 2–3 years / Unlikely within 3 years]*

Appendix B. Detailed Sample Characteristics

This appendix provides the full demographic profile of the survey sample (N = 136). All respondents met the inclusion criteria: direct involvement in AI strategy or implementation at organizations with 50 or more employees.

TABLE III. DISTRIBUTION BY INDUSTRY SECTOR (N = 136)

Industry Sector	n	%
Technology	37	27.2
Financial Services	26	19.1
Manufacturing	19	14.0
Retail & E-commerce	15	11.0
Healthcare & Pharmaceuticals	7	5.1
Telecommunications	6	4.4
Energy & Utilities	5	3.7
Education	5	3.7
Government & Public Sector	4	2.9
Professional Services & Consulting	4	2.9
Transportation & Logistics	3	2.2
Media & Entertainment	2	1.5
Real Estate & Construction	2	1.5
Other	1	0.7
Total	**136**	**100.0**

TABLE IV. DISTRIBUTION BY RESPONDENT ROLE (N = 136)

Role	n	%
Product Manager	35	25.7
Engineering Lead / Team Lead	30	22.1
CTO / VP of Technology	23	16.9
AI/ML Engineer	22	16.2
Business Analyst	16	11.8
Data Scientist	5	3.7
Other (IT Director, Architect, etc.)	5	3.7
Total	**136**	**100.0**

TABLE V. DISTRIBUTION BY ORGANIZATION SIZE (N = 136)

Organization Size	n	%
< 500 employees	40	29.4
500–5,000 employees	46	33.8
> 5,000 employees	50	36.8
Total	**136**	**100.0**

TABLE VI. GEOGRAPHIC DISTRIBUTION (N = 136)

Country	n	%
Russia	48	35.3
United States	20	14.7
United Kingdom	14	10.3
Germany	11	8.1
India	11	8.1
United Arab Emirates	8	5.9
Netherlands	7	5.1
Singapore	7	5.1
Brazil	10	7.4
Total	**136**	**100.0**

TABLE VII. SELF-ASSESSED EAAMM MATURITY LEVEL (N = 136)

EAAMM Level	n	%
L1 — Assistive	56	41.2
L2 — Collaborative	43	31.6
L3 — Consultative	26	19.1
L4 — Supervisory	8	5.9
L5 — Autonomous	3	2.2
Total	**136**	**100.0**

TABLE VIII. AI GOVERNANCE FRAMEWORK STATUS (N = 136)

Governance Status	n	%
Formal AI governance framework	24	17.6
Partial framework	47	34.6
No framework	65	47.8
Total	**136**	**100.0**

Enhancing Employee Innovation and Organizational Agility: Using Transformational Leadership in UAE Smart Government Entities

Khalfan Mohammed Alshamsi
Abu dhabi School of management
adsm-215873@adsm.ac.ae
ORCID: 0009-0002-4415-6085

Abualla Ali Alketbi
Abu dhabi School of management
adsm-215871@adsm.ac.ae
ORCID: 0009-0004-4228-9719

Dr. Turki Al Masaeid
Abu dhabi School of management
t.almasaeid@adsm.ac.ae
ORCID: 0000-0002-7901-5656

ABSTRACT— The study examines how transformational leadership can contribute to the innovation and agility of employees and organizations in UAE smart government institutions. Even though a lot of money has been spent on the digital infrastructure most government departments have not been able to develop innovative organizational cultures that can harness all the technological abilities. Using a quantitative survey that was conducted with 150 government employees that work in three large entities in the UAE, this paper was able to investigate the effects of transformational leadership in enhancing the innovative behaviors of employees and responsiveness of organizations. The results indicate that transformational leadership has a strong influence on the worker innovation ($\beta = 0.68$, $p < 0.001$) and organizational nimbleness ($\beta = 0.72$, $p < 0.001$). The main suggestions are to create the leadership programs based on transformational skills, develop the innovation systems, develop the cross-functional digital teams, and introduce the performance measurements in accordance with the innovation outcomes. The present study would be relevant to the comprehension of leadership-powered digital transformation in the government setting and offer practical suggestions to UAE policymakers that are working on the Digital Government Strategy 2025.

Keywords—Transformational Leadership, Employee Innovation, Organizational Agility

I. INTRODUCTION

The UAE has placed itself among top world leaders in digital transformation of the government due to the UAE Digital Government Strategy 2025. The overall framework is new to bring about full digitization of the government services with the focus being to reach 90% federal services completed end to end digitally. The plan focuses on digital by design, data-centric strategies, being proactive and delivering services to people in a user-centric manner with a goal of training 10,000 federal government workforces on the use of artificial intelligence, blockchain and Internet of Things technologies.

Nevertheless, even with the significant investment in technology a significant disparity exists between the digital infrastructure implementation and the innovation capabilities of organizations. The challenge is that many of the UAE government departments have sophisticated digital resources but cannot create an organizational culture that is creative enough to utilize them to the fullest. It is no longer a technological infrastructure that is needed in the digital transformation process but leadership strategies that ensure employees are creative, take calculated risks, and respond to changing needs of citizens quickly.

The UAE smart government organizations are in a paradox situation wherein massive digital infrastructure investments have not been proportionately transferred to the creation of innovative organizational behaviors and responsive operations. Digital technologies are usually implemented by government departments in a reactive manner when a particular tool is used to replicate an already-existing process instead of entirely reinventing the service delivery paradigm. Workers can have access to a sophisticated system without the psychological safety, motivation, and leadership that support experimenting with new ways of doing things respectively trying out old ones.

Moreover, organizational agility the ability to feel environmental shifts and react quickly by reconfiguring resources and processes has not been well developed in most government organizations. The nature of traditional public sector such as risk aversion, rigidity of procedures and hierarchical decision-making systems prevents the ability to take flexible and adaptive behaviors that are imperative in governance in digital age.

The proposed research is expected to investigate the application of the transformational leadership concept in improving employee innovation and organizational agility in smart government organizations in the UAE. The purpose of the specific objectives is to: (1) understand the exercise of current transformational leadership practices and scope of gaps impeding innovation; (2) assess the relationship between transformational leadership actions and employee behaviors indicating innovative work to the organization; (3) determine the influence of transformational leadership on organizational agility; (4) identify the mediator variables e.g. organizational culture which has been facilitating or restricting the effects of transformational leaders; as well as (5) develop actionable recommendations to government leaders and policymakers.

II. LITERATURE REVIEW

A. Transformational Leadership Theory

Transformational leadership: Transformational leadership was developed by Bass and Avolio as a paradigm in which leaders inspire followers to be more motivated, morally uplifted, and working more effectively by their ability to articulate a vision and through inspirational communications and providing a good example. The model considers four dimensions which include the idealized influence (role modelling), inspirational motivation (communication of strong visions), intellectual stimulation (fostering creativity and challenging assumptions), and an individualized consideration (attention to development needs) dimension. Studies prove that the ability of transformational leadership to create innovation-fearing conditions through the development of psychological safety allows employees to suggest innovative ideas without the fear of being penalized (Afsar and Umrani, 2020).

B. Employee Innovation in Digital Transformation

The innovative behavior of employees includes identifying the problem, developing innovative ideas, promoting, and putting new ideas to practice. In digital transformation situations, innovation means not just technology adoption, but process redesigning, service model innovation and cooperation with a focus on harnessing the digital potential. The studies conducted in the framework of the UAE public sector confirm that HRM practices, innovation climate, and the productivity of the employees proved to be highly interrelated (AlDhaheri et al., 2023). Research indicated that difficult appraisal-framing digital transformation as an opportunity and not a threat and organizational culture support play a very important role as critical mediators between digital transformation initiatives and innovative behaviors of employees.

C. Organizational Agility and Digital Transformation

Organizational agility is the ability to perceive the changes in the environment, make fast decisions and respond to changes by reconfiguring resources and adapting processes. It has been shown that organizational agility is a mediator variable that plays a critical role in linking digital transformational leadership and desirable results of digital transformation (AlNuaimi et al., 2022). Research based on structural equation modeling shows that digital transformational leadership is a significant contributor of organizational agility that contribute to digital transformation success. This relationship under mediation suggests that leadership will have effects on digital transformation results in the form of creating organizational abilities to adapt quickly, as opposed to technological intervention.

D. UAE Smart Government Context

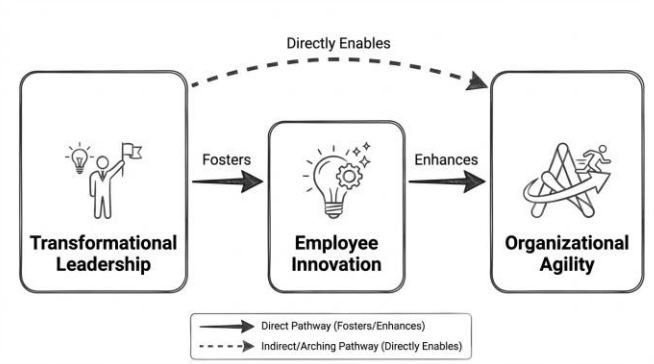
The Digital Government Strategy 2025 of the UAE lays up an elaborate set of structures which include digital enablers such as unified digital identity schemes, national digital wallets, and national online signature. Nevertheless, studies that analyze the topic of UAE government digital transformation indicate that certain challenges are still present. In the research conducted by Al Qudah (2024) on 223 government employees, digital transformation adoption has challenges that have to be overcome with prudent leadership methods that focus on the quality of decisions, the effectiveness of communication, and ethical issues. The research examining the smart government effectiveness in the UAE public organizations illustrates that internal government efficiency and infrastructure to implement the smart government services are significant predictors of smart government success (Abdulqader and Al Marri, 2020). Based on the existing literature, this paper presents a theoretical framework that examines the connections between transformational leadership, employee innovation, and organizational agility in smart government organizations in the UAE. Transformational leadership should enhance innovation among the employees through intellectual stimulation and psychological empowerment. Moreover, it is bound to have a direct influence on the level of organizational agility as it facilitates quicker decision-making and responsive actions. Organizational agility is also expected to improve through employee innovation that would facilitate the execution of emerging ideas and responsive practices. Based on that, the current study develop its hypothesis and conceptual model:

H1: Transformational leadership has a positive impact on innovation by employees.

H 2: Transformational leadership has positive effects on organizational agility.

H3: Organizational agility is positively affected by employee innovation.

Figure 1: conceptual model



III. DATA METHODOLOGY

A. Research Design

Using a quantitative approach, the proposed research involves the use of a structured questionnaire survey as a study instrument to identify correlation between transformational leadership, employee innovation, and organizational agility. The cross-sectional survey design will allow the systematic measurement of leadership behaviors, the perceptions of employees and organizational

outcomes as well as the statistical analysis of relationships strength and pattern.

B. Sample and Data Collection

The target market will include employees of the federal and emirate government agencies located in the UAE who are currently involved in smart government programs. A purposive sampling method was used to identify government organizations that have prepared digital transformation programs. The sample contained workers in three large state organizations who were representing various functional areas. One hundred and fifty complete responses were received in online surveys distribution in a four week time period, a response rate of 62 percent.

C. Measurement Instruments

Transformational leadership was assessed with the help of the modified Multifactor Leadership Questionnaire which was based on four dimensions. The innovative behavior of employees was evaluated using items assessing the idea generation behavior, promotion behavior, and implementation behaviors that concerned the nature of digital transformations. The measurement of organizational agility was made through the items that evaluated the sensing abilities, speed, and flexibility of decisions, and implementation. The scoring scales were all Likert five-point scales. The demographic questions included in the questionnaire gathered information on age, sex, education, tenure in the organization and level of position.

D. Data Analysis

Data screening, followed by calculating descriptive statistics, reliability, and correlation analyses were performed to examine bivariate relationships and determine primary hypotheses of the research taking into account the demographics. Further discussion involved possible mediating variables through which transformational leadership can have an influence.

E. Measurement Model Assessment

The constructs were tested with the help of Confirmatory Factor Analysis (CFA) to determine the validity and reliability of the measures. Measurement model was found to have acceptable fit indices, with the indices being greater than recommended values (CFI > 0.90, TLI > 0.90, RMSEA < 0.08, SRMR < 0.08).
Convergent validity was achieved since all the loading factors were significant and greater than 0.60. Also, the all constructs values of AVE were greater than 0.50 and the values of Composite Reliability (CR) were greater than 0.70; these values were satisfactory internal consistency values.

F. Ethical Considerations and Limitations

The study was conducted according to ethical principles such as informed consent, protection of confidentiality and voluntary participation. Permission to conduct the surveys was received by the organizations in advance. Limitations involve that cross-sectional design did not allow attributing definite causality, the possible risk of self-report bias due to the use of self-report questions, purposive sampling, and possible generalizability as it was narrow to UAE contexts, and the possibility that there was a limit in the scope of transferability to other contexts due to the cultural circumstances of the UAE. Due to the self-reported information, a single factor test of Harman was performed to determine whether common method bias was present. These findings were that there is no one factor that explained the most variance (less than 50%), which means that the common method bias is not a major issue in this research.

IV. FINDINGS, ANALYSIS, AND DISCUSSION

A. Sample Characteristi*cs*

The population sample so formed was 150 government employees containing 58 and 42 per cent males and females respectively. The age range was 23% at the age of 25-30, 31% in the age range of 31-35 years, 28% in the 36-40 years age range, and 18% above 40 years. The level of education was found to be highly qualified with 67% having bachelor degrees, 28% with master degrees while 5% had doctoral qualifications. The position levels were 41 percent frontline, 43 percent supervisory staff as well as 16 percent middle management.

B. Descriptive Statistics and Reliability

Transformational leadership had a mean rating of 3.62 (SD = 0.74) on the five-point scale, employee innovative behavior had a mean of 3.48 (SD = 0.68), and organizational agility had a mean of 3.71 (SD = 0.62). Internal consistency was found to be acceptable according to the reliability analysis: the transformational leadership had a Cronbach alpha of 0.89, employee innovative behavior of 0.86 and organizational agility of 0.88.

Table 1: Descriptive Statistics and Reliability Coefficients

Variable	**Mean**	**SD**	**Cronbach α**
Transformational Leadership	3.62	0.74	0.89
Employee Innovative Behavior	3.48	0.68	0.86
Organizational Agility	3.71	0.62	0.88

C. Correlation Analysis

The Pearson correlation analysis showed that there were significant positive correlations between the primary study variables. Transformational leadership showed a good positive relationship with employee innovative behavior ($r = 0.65$, $p < 0.001$) and organizational agility ($r = 0.69$, $p < 0.001$). Innovative behavior among the employees and agility of an organization were highly correlated ($r = 0.58$, $p < 0.001$). Looking at transformational leadership dimensions in isolation, intellectual stimulation was found to be having the highest correlations with both the innovative behavior ($r = 0.71$, $p < 0.001$) and the organizational agility ($r = 0.68$, $p < 0.001$).

Table 3: Correlation Matrix (*** $p < 0.001$)

Variable	TL	EIB	OA
Transformational Leadership (TL)	1.00	-	-
Employee Innovative Behavior (EIB)	0.65***	1.00	-
Organizational Agility (OA)	0.69***	0.58***	1.00

D. Regression Analysis

Various regression analysis on the effect of transformational leadership on employee innovative behavior was statistically significant ($F = 45.23$, $p < 0.001$) with a variance explained amounting to 48 ($R^2 = 0.48$). The strongest predictor ($\beta = 0.68$, $t = 9.34$, $p < 0.001$) was the transformational leadership. Education level is the only demographic control that had a significant relationship with innovative behavior ($\beta = 0.18$, $p < 0.05$). Additional result showed that when all the dimensions were simultaneously tested, intellectual stimulation ($\beta = 0.52$, $p < 0.001$) and individualized consideration ($\beta = 0.31$, $p < 0.01$) were significant to predict innovative behavior.

Table 3: Regression Analysis Results (*** $p < 0.001$, * $p < 0.05$, ns = not significant)

Predictor	Employee Innovation β	Organizational Agility β
Transformational Leadership	0.68***	0.72***
Education Level	0.18*	0.12 ns
R^2	0.48	0.53

Regression analysis of the organizational agility was statistically significant ($F = 52.18$, $p = 0.001$) and explained 53 percent of variance ($R^2 = 0.53$). Transformational leadership was found to be a strong predictor of agility in organizations ($\beta = 0.72$, $t = 10.67$, $p < 0.001$). The impact of demographic variables on organizational agility perceptions was low. Idealized influence ($b = 0.38$, p under 0.01) and the intellectual stimulation ($\beta = 0.45$, p under 0.001) were the strongest predictors in simultaneous regression.

Figure 1: Transformational Leadership and Employee Innovative Behavior

Figure 2: Transformational Leadership and Organizational Agility

E. Key Survey Findings

Survey responses provided additional nuanced insights. Most of the respondents (68 percent) were of the opinion that their leaders do push them to look beyond the old frameworks, but only half of the respondents thought that their leaders actively solicit their ideas on how to do services better. To measure innovative behaviors, 71 percent of the respondents claimed to be frequently able to find willingness to enhance service delivery with digital tools, however, 43 percent commonly use innovative ideas. This is an implementation gap that focuses on an obstacle between the responsiveness of the idea and its execution. Also, a third of them were worried of the negative implications of failed innovation efforts indicating that risk aversion is still a major innovation barring factor.

Mixed perceptions were found with organizational agility items. Whereas 67% considered that their organizations are responsive when it comes to responding to the feedback of the citizens, 49% of the respondents felt that their organizations were quick to embrace emerging new technologies whenever they came across such opportunities. In addition, half expressed that bureaucracy is an issue that slows down implementation and forty two percent believed the process of decision making was too centralized.

F. DISCUSSION

The results offer solid empirical evidence in support of the transformational leadership importance in promoting employee innovation and organizational agility in entities of smart government in the UAE. The positive close relationships that were observed are consistent with theoretical propositions and complement the body of current knowledge as they indicate that these relationships are relevant in the context of digital transformation in the public sector of UAE. The fact that intellectual stimulation has such a potent effect on all the outcomes deserves a mention, and the focus on preventing simplistic assumptions that is inherent in this dimension seems particularly appropriate in the context of digital transformation that demands reconsidering the core processes.

The lag between idea generation and implementation points to a serious issue that needs particular consideration. Whereas transformational leaders are effective in generating creative thinking, structural and systemic obstacles hinder the translation of ideas into solution implemented. It implies that wholesome digital transformation should be taken up concurrently whereby leadership development, organizational structure redesign, resource allocation schemes, and performance evaluation systems that reward innovation efforts irrespective of the short-term results are concerned. Besides the identified direct relationships, the findings support the theoretical assumption that leadership behaviors are central to supporting innovation-driven agility. Transformational leadership, especially via intellectual stimulation, seems to establish a climate of experimentation and adaptability of thought. This is particularly important in the field of digital government, where technology evolves quickly, necessitating constant learning and responsiveness.

V. RECOMMENDATIONS

A. Leadership Development Initiatives

The UAE government agencies would need to institute rigorous transformational leadership development initiatives focused on the leadership of all organizational levels in terms of intellectual stimulation and personalized consideration skills. The training modules can incorporate real-life activities of getting to challenge assumptions, conducting creative problem-solving workshops, and offer individual coaching. The introduction of digital transformation-specific content on the issues of application of the transformational behaviors in the environment of the fast change of technology should be introduced into the leadership development. Having mentoring programs in which experienced transformational leaders are paired with the emerging leaders can be used to transfer skills practically.

B. Innovation Infrastructure and Support Systems

To overcome the implementation gap, the state actors are supposed to have formal innovation infrastructure that helps in idea pushing through idea conception stages to implementation. This infrastructure must have specific innovation teams that review employee proposals, rapid prototyping capabilities that can be used to run tests quickly, and simplified processes that can help facilitate innovation pilot. The mechanisms used to allocate resources should be rigorous in the integration of innovation support, whereby resources available to employees in terms of time, funding and technical support are known. Success and failure Performance appraisal systems need to be reformulated to reward any innovation efforts which should be explicit that not all innovations would be successful.

C. Organizational Structure and Process Redesign

Government agencies need to establish and remove procedural processes that are now redundant in digital operation space in terms of critical control and accountability roles. Decisions relating to the company also ought to be delegated to the relevant levels such that frontline employees and the supervisors make swift decisions without having a long approval chain. The digital transformation teams that cross-functional team comprised of employees representing various departments would increase the levels of agility and innovation through knowledge sharing and silos reduction as well as ensuring a quicker response to mobilize the capabilities that existed across organizational boundaries.

D. Culture Change and Performance Metrics

The root cause of continuing risk aversion must be tackled through culturally planned efforts to stress that risk-taking based on calculations and experience-based learning are important elements that should be appreciated. Top management members must repeatedly convey the idea that innovation is a risk and that companies know what to do when the initiative fails. Government agencies are advised to put in place elaborate measures that monitor the innovation efforts, performance agility and transformational leadership practice. Continuous review and reporting of such metrics to the top management keeps the priorities of innovation and nimbleness on the strategic agenda with other conventional performance and compliance indexes.

According to the empirical data, a number of specific recommendations can be suggested:

First, due to the high positive relationship between transformational leadership and employee innovation ($b = 0.68$), leadership development interventions in government organizations in the UAE should focus on the intellectual stimulation and individualized consideration competencies. These dimensions proved to be the most predictive and are important in encouraging innovative behaviors amongst the employees.

Second, given that the concept of transformational leadership and that of organizational agility are strongly interrelated ($b = 0.72$), the organizations are supposed to make institutional leadership practices that favor decentralized decision-making and flexible governance systems. This will increase responsiveness and decrease delays of bureaucracies.

Third, the fact that employee innovation and organizational agility have a positive relationship ($r = 0.58$) demonstrates the necessity to reinforce the innovation implementation mechanisms. The governmental

organizations are to introduce the systematic pipelines of innovation (i.e., idea screening, rapid prototyping, cross-functional collaboration platforms).

Lastly, to bridge the perceived gap between the idea generation and implementation, the performance management systems are to be restructured to incentivize the innovation efforts, even failed efforts to prevent risk aversion and promote constant experimentations.

VI. CONCLUSION

Empirical data on transformational leadership revealed in this research indicates that such leadership is of paramount importance in promoting innovation and agility of employees in smart government organizations in the UAE. Even though the investment in digital infrastructure is significant, to achieve all the potential of transformation it is important to find leadership styles that are particularly used in developing new behaviors of employees and responsiveness of organizations. Intellectual stimulation and one-on-one consideration dimensions of transformational leadership are influential in forecasting employee innovative working behaviors and perceptions of agility in an organization among employees working within the UAE government.

Results show that although UAE government agencies have realized some advancement in digital transformation, there are a series of challenges such as risk aversion, implementation barriers, and structural constriction of agility that still restrain the realization of innovations. The following issues should be systematically addressed by means of leadership development efforts, establishment of an innovation infrastructure, the redesign of the organization, culture change efforts, and the proper performance measurement. Effective digital change does not merely consist of the deployment of technology but poses inherent requirements of the human and organizational capacity development.

The process of full achievement of smart government in the UAE will be defined not only by the level of technological development but also leadership, organizational flexibility, and innovation potential of employees. Transformational leadership is an established method of driving the human and cultural change required to support digital initiatives to reach their scalp in improving citizen services, efficiency in operations, and competitive advantage on the national level. Transformational leadership development should be among the key strategic investments by the government leaders and policymakers in the exploration of agile returns in the long term in the form of innovation, agility, and eventually, the fulfilment of digital government visions.

This research paper will add value to the body of existing literature by empirically confirming the connections between transformational leadership, employee innovation, and organizational agility in the framework of UAE smart government organizations. It builds on the previous studies by illustrating that the influence of behavioral aspects of leadership is fundamental in the transformation of digital infrastructure investments into concrete innovation and agility results.

This study has a number of limitations despite its contribution. The cross-sectional design does not allow causal inference, and future research should consider longitudinal approaches which provide a better capture of dynamic relationship in the course of time. Also, it is recommended that future researchers should use structural equation modeling to better test mediation effects and increase the rigor of analysis.

REFERENCES

[1] Afsar, B., & Umrani, W. A. (2020). Transformational leadership and innovative work behavior: The role of motivation to learn, task complexity and innovation climate. European Journal of Innovation Management, 23(3), 402-428. https://doi.org/10.1108/EJIM-12-2018-0257

[2] AlDhaheri, H., Hilmi, M. F., Abudaqa, A., Alzahmi, R. A., & Ahmed, G. (2023). The relationship between HRM practices, innovation, and employee productivity in UAE public sector: A structural equation modelling approach. International Journal of Process Management and Benchmarking, 13(2), 157-178. https://doi.org/10.1504/IJPMB.2023.128471

[3] AlNuaimi, B. K., Singh, S. K., Ren, S., Budhwar, P., & Vorobyev, D. (2022). Mastering digital transformation: The nexus between leadership, agility, and digital strategy. Journal of Business Research, 145, 636-648. https://doi.org/10.1016/j.jbusres.2022.03.038

[4] Al Qudah, M. (2024). Digital transformation and wise leadership: Challenges and opportunities evidence from government institutions in UAE. Journal of Police and Legal Sciences, 15(2), Article 5. https://doi.org/10.69672/3007-3529.1026

[5] Abdulqader, A., & Al Marri, K. (2020). The influence of transformational leadership style on innovation behaviours: The case of the government sector of the UAE. In A. Abu-Tair, A. Lahrech, K. Al Marri, & B. Abu-Hijleh (Eds.), Proceedings of the II International Triple Helix Summit (pp. 3-16). Springer. https://doi.org/10.1007/978-3-030-23898-8_1

[6] Alakaş, E. Ö. (2024). Digital transformational leadership and organizational agility in digital transformation: Structural equation modelling of the moderating effects of digital culture and digital strategy. International Journal of Information Management Data Insights, 4(2), Article 100269. https://doi.org/10.1016/j.jjimei.2024.100269

[7] Ly, B. (2024). The interplay of digital transformational leadership, organizational agility, and digital transformation. Journal of the Knowledge Economy, 15(1), 985-1015. https://doi.org/10.1007/s13132-023-01377-8

[8] Mergel, I., Ganapati, S., & Whitford, A. B. (2021). Agile: A new way of governing. Public Administration Review, 81(1), 161-165. https://doi.org/10.1111/puar.13202

[9] Muduli, A., & Choudhury, A. (2024). Exploring the role of workforce agility on digital transformation: A systematic literature review. Benchmarking: An International Journal, 31(4), 1285-1310. https://doi.org/10.1108/BIJ-03-2023-0156

[10] Shehadeh, M., Almohtaseb, A., Aldehayyat, J., & Abu-AlSondos, I. A. (2023). Digital transformation and competitive advantage in the service sector: A moderated-mediation model. Sustainability, 15(3), 2077. https://doi.org/10.3390/su15032077

[11] Trischler, J., & Westman, T. J. (2021). Design for experience: A public service design approach in the age of digitalization. Public Management Review, 24(8), 1251-1270. https://doi.org/10.1080/14719037.2021.1899272

[12] United Arab Emirates Government. (2024). UAE Digital Government Strategy 2025. Retrieved from https://u.ae/en/about-the-uae/strategies-initiatives-and-awards/strategies-plans-and-visions/government-services-and-digital-transformation/uae-national-digital-government-strategy

[13] Weber, E., Büttgen, M., & Bartsch, S. (2022). How to take employees on the digital transformation journey: An experimental study on complementary leadership behaviors in managing organizational change. Journal of Business Research, 143, 225-238. https://doi.org/10.1016/j.jbusres.2022.01.036

[14] Zafar, H., Ko, M. S., & Osman, I. H. (2024). Navigating digital transformation in the UAE: Benefits, challenges, and future directions in the public sector. Information, 13(11), 281. https://doi.org/10.3390/info13110281
[15] Zhang, M., Chen, X., Xie, H., Esposito, L., Parziale, A., Taneja, S., & Siraj, A. (2024). Top of tide: Nexus between organization agility, digital capability and top management support in SME digital transformation. Heliyon, 10(10), e31579. https://doi.org/10.1016/j.heliyon.2024.e31579

Determining the Optimal Drilling Program in Oil & Gas using modern AI (Conceptual Framework)

Maryam AlShehhi
Student Master Business & Analytics
Abu Dhabi School of Management
Abu Dhabi, UAE
Maryam_saleh@live.com

Dr. Ahmad Jaffar
Associate Professor of Information Systems
Abu Dhabi School of Management
Abu Dhabi, UAE
a.jaffar@adsm.ac.ae

Abstract (Preparing a drilling program is a critical stage in oil and gas well delivery, where engineers must define operational parameters, anticipate geological conditions, and establish appropriate risk control measures before drilling operations begin. Traditionally, this process relies heavily on manual analysis of offset wells, historical reports, and individual engineering experience. While effective, such practices are time-consuming and may lead to inconsistencies in planning quality, incomplete knowledge transfer from previous wells, and repeated operational issues such as non-productive time (NPT), fluid losses, or stuck pipe incidents.

This research proposes an AI-enabled conceptual framework for drilling program preparation that transforms conventional manual workflows into a structured, data-driven planning process. The framework integrates hierarchical offset well selection, machine learning analysis, and rules-based engineering governance to support systematic extraction of historical drilling parameters and operational risks. Relevant offset wells are automatically identified based on reservoir similarity and spatial proximity, after which key planning parameters and historical drilling events are analysed using machine learning models to generate predictive insights and recommended operational envelopes. A rules-based governance layer ensures compliance with engineering standards, regulatory constraints, and safety requirements while maintaining traceability of decision logic.

The proposed framework functions as a decision-support system designed to assist drilling engineers in developing consistent, risk-aware drilling programs while preserving human oversight and professional judgment. The study contributes to advancing digital well planning by demonstrating how artificial intelligence can enhance efficiency, knowledge transfer, and planning reliability in complex drilling environments.)

Keywords: **Artificial Intelligence in Drilling, AI-Driven Well Planning, Offset Well Analysis, Machine Learning Optimization, Rules-Based Governance, Risk-Aware Drilling Programs.**

I. INTRODUCTION

Preparation of a drilling program is a preparative step in well delivery (Al-Mudhafar, 2023), which entails engineers or deciding on parameters of action, formation behavior and defining risk controls before action. This practically relies heavily on the review of the historical data on drilling, engineering guidelines and technical judgment. As the drilling effect increases in both mature and complex reservoirs so does the quantity of available well data that complicates the extraction of key learning in a rapid and reliable fashion. Most of the planning processes nowadays are supplemented by hand search in the historic wells, report analysis and adapting the learning to new objectives. This, though efficient, method consumes time and is based on personal experience which predisposes various QA/QC procedures with the aim of eliminating any variation in the quality of the programs. An organised way of storing the experience of the past and presenting it in the form that can be reused would improve the uniformity in the planning and assist the engineer to apply the best practice in wells that share similar geological and operating history (Mohammadinia, et al. 2025).

The increasing capabilities of digital technology and data-driven techniques provide an occasion to improve the early stage well planning by structuring and interpreting historical drilling data in a uniform way. It will allow engineers to gain access to pertinent insight into operations more effectively, enhancing the quality of the decisions and minimizing the chances of overlooking essential risks.

II. PROBLEM STATEMENT

The construction of drilling programs is still an art based on reading the offset wells and relying on the interpretation of the engineer resulting in uneven plans and repetitive lack of lessons learned on the neighboring wells (Petrofac, 2024). None of the systems will automatically identify the appropriate offset wells regarding positions and reservoir will recover the planning logic and document the issues like losses, fractures or stuck pipe and generate a drilling program that displays such findings (Amani, et al., 2024). As such, known risks become re-discovered, instead of being avoided and planning quality is built more on the experience of the individual engineer on the experience of earlier wells.

III. AIM, OBJECTIVES AND SCOPE

Aim

The thesis aims to develop conceptual framework and test an AI based drilling program preparation that systematically uses past data on offset wells, machine learning expertise and rule based governance to increase consistency, traceability and risk awareness during the stage of planning the delivery of oil and gas wells.

Objectives

To achieve this aim, the thesis will attempt to achieve the following objectives:

1. To study the structure of the preparation of the drilling programs and the existing industry practice. It entails the examination of the existing approaches to determining the offset wells, extract planning parameters, and transferring the historical risks, which are now determined by drilling engineers, and checking the existing digital and AI-supported planning approaches in literature and practice.

2. To understand and identify the key factors that need to be incorporated in an AI-based drilling planning model: Evaluate current AI- technology which can be streamlined in the drilling planning system.

These elements are hierarchical offset-well selection logic, historical parameter and risk extraction mechanisms, appropriate machine-learning model roles (prediction and optimization), and rule-based elements of governance that are required to meet safety, standardization, and traceability.

3. To prepare a conceptual framework of the drilling program preparation: The framework brings together the proposed elements as an end-to-end planning process that depicts the interaction of the offset-well data, AI models, and engineering rules to create a consistent and risk-aware drilling program and retain human control.

4. To test the framework created with expert review and qualitative validation: The clarity, practicability, completeness and the capability of the framework to improve the quality of the planning are assessed by structured feedbacks of drilling and well-planning professionals in comparison with the traditional manual approaches.

Scope

The study is only limited to the well planning stage of drilling operations and does not include real time optimization of drilling, automated control and field implementation. It is an abstract and procedural structure, it defines the logic of the workflow, relations of data and model functions, yet it does not include software code or field implementation.

The analysis assumes that the records of the historical offset-wells of the same reservoir or a particular radius of space are available and the company standards, regulatory and safety restrictions are assumed to be the hard determinants that are required to be incorporated in the model. It is aimed at aiding and standardisation of the engineering decision-making process, not to exclude professional judgment and to hold people responsible.

IV. METHODOLOGY

The chapter gives the research procedure that will be applied in the preparation and testing of the proposed AI conceptual framework of drilling program preparation. This chapter was to address the research design, the procedure of data collection and data analysis procedures. This research is a theoretical research which does not presuppose the creation of software and the introduction of field systems, hence, the methodology is developed to extract the knowledge existing in the literature and verify the framework suggested by the researcher in the process of systematic cooperation with the experts. The other ethical concern, which is considered in the chapter, is the ethical considerations that are directed to inform the research process ensuring that all professional contributions are handled in a responsible and professional way that meets the needs of the institution and industry.

The overall approach is a qualitative research paradigm, which is appropriate in a study that emphasises exploration, interpolation, and explanation of a complex phenomenon but not quantifying and measuring it (Creswell and Creswell, 2018). Of interest to the context of the present research is the way drilling engineers develop well programs basing on historic data about offset wells and to what extent a systematic, AI-aided framework is able to contribute to the uniformity, traceability and risk awareness of the process. It adopts a secondary research method as the key means of data collection and the thematic analysis using NVivo *figure4* software is taken. Such a combination allows the scholar to conduct a systematic contribution according to the content of the existing academic and industry literature, outline common patterns and themes, and implement the findings on the formation of the conceptual framework.

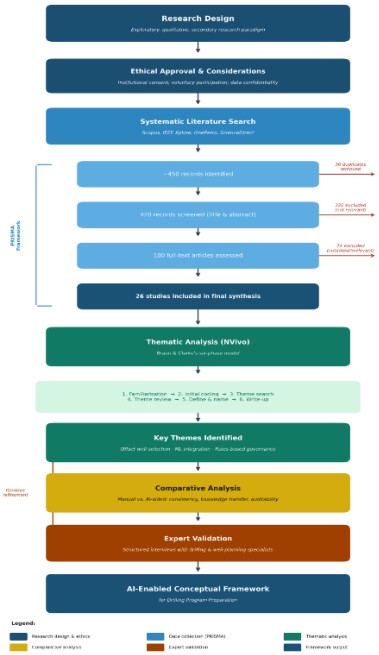


Figure 3 The PRISMA-guided literature selection procedure is represented in the left bracket and the amount of exclusion at each phase. Colour coding separates the research design (dark navy), data collection (blue) and thematic analysis (green), the comparative

A. Ethics in Research

The importance of ethics in any research work is also important in a study on human subject directly or on secondary literature. Ethical consent has also been taken in this study and agreed upon prior to the involvement of any industry practitioners or taking any form of feedback. The step will ensure that the research satisfies both the institutional ethical standard and an interest to carry out research in a responsible manner (Bryman, 2016).

Although the primary data collection in the specified research will primarily rely on the secondary one, encompassing journal articles, conference papers, and industry reports published, the validation stage of the framework will presuppose the structured interview with the drilling and well-planning specialists. During this stage, the suggested framework will be presented to the participants in order to provide feedback on its clarity, feasibility, and relevance to the real process of planning. To protect the rights and the interests of such participants, there are different ethical in place. One, it is entirely voluntary and the participants may leave any time without one stating the reason. Second, they receive informed consent before any set of feedback is given and the aim of the study is made clear, the data is being used is also made clear and measures to guarantee the confidentiality of the information are made (Saunders, Lewis, and Thornhill, 2019). Third, no personal data, proprietary well identifiers and commercially sensitive data are obtained during the consultation. The interaction is strictly confined to the framework, reason, and viability of the proposed framework against a specific operational data of a company or an individual.

Other than guarding the participants, the study is also ethical in regard to conscionable use of second-hand information. All sources that were used in the literature review have been cited and referred to as per the academic standards. All the information is neither produced nor modified to support the results and the evaluation is conducted in an open manner that allows having an independent verification. Besides, the proposed framework is presented as a decision-support system, but not a decision-making system. This distinctiveness is very important in that it ensures that the professional responsibility and engineering judgment will not be delegated to the framework and any automated process but the human engineers. This is a moral stand that is maintained throughout the process of conducting the research and all the interaction processes captured in the documentation.

B. The Methodology

1.) Description of research Design

The research design used in the study is the qualitative and exploratory research design. Exploratory design is applicable in the case when the research objective is to explore the issue, which is not completely covered in the literature and define a new meaning or a model based on the available evidence (Saunders, M., Lewis, P., and Thornhill, A. (2019)). However, based on what is established in the literature review, much of the aspects of machine learning to predict the parameters of drilling, offset well analysis and rule-based governance have been considered in isolation, but no current literature has offered a systematic end-to-end process of implementing all these in a single process of planning. This ambiguity justifies an exploratory methodology that tries to create such a framework at its fundamental basis basing on the trends and findings suggested in the published works.

The research is based on both secondary and primary data. The secondary data is provided by academic journals, conference papers, and industry reports that are published. The primary data is collected amongst drilling and digital professionals through structured interviews and questionnaire survey. The combination of the two kinds of data makes the study more solid, since the secondary data give the theoretical background, whereas the primary data help to verify the

practicality and usefulness of the framework in actual operations (Creswell, J. W., and Creswell, J. D.). (2018)).

2.) Data Collection Methods

Secondary Data

The information utilized in this study is collected on the form of systematic and structured review of the available scholarly and industrial literature. The search strategy is based on the PRISMA framework that offers a clear and repeatable way of finding, filtering, and choosing the sufficient studies (Page, M. J., et al. (2021)). The databases utilized are the Scopus, IEEE Xplore, the OnePetro and the ScienceDirect databases which are the most notable databases to conduct research in engineering and technology within the oil and gas industry. Some of the search terms include the AI in well planning, digital well planning, offset well analysis, machine learning drilling parameters, Bayesian optimisation drilling, drilling risk transfer and rules-based systems engineering, which are the key word phrases.

The query on these databases gave out about 450 records. After deleting 30 of the records that were noted to be duplicates, the rest of the records of 420 were narrowed down in regards to their titles and abstracts in order to ascertain the relevance of the records to the research questions. Such screening allowed cutting the 300 records to 100 that were further narrowed down through reading their contents more closely. The last category of the primary synthesis is a total number of 26 studies that formed the foundation of evidence of the literature review and the development of the conceptual framework (Vishnumolakala, N. (2022)).

The collection of the data was achieved by examining the industry reports, the technical guidelines and white papers of the large oil and gas operators and service companies besides the journal articles and the conference articles. These sources were rather convenient in providing the practical background and real life examples to the academic literature and in ensuring that the framework under consideration is grounded in the practical reality rather than theoretical literature (Petrofac.). (2024)).

Primary Data: Expert Interviews

The proposed conceptual framework was to be proven by conducting structured interviews with industry professionals. The sample population included the drilling engineers and digital or AI professionals who were directly engaged in the planning of wells and in making the operational decisions. The interviews were structured i.e. all the interviewees were asked the same set of questions and in the same order. The responses can be compared to the participants and the strategy is similar (Bryman, A.). (2016)).

The interviews were also focused: they had to obtain the expert feedback on the proposed AI-based drilling framework. The participants were taken through the framework diagram and requested to comment on the design, logic and usefulness in practice. Two sets of professionals were identified in order to find out the rationale behind the decision of offsets, risk transfer procedure, and planning output are pragmatic, reflect the actual operational needs and digital or AI professionals, to find out whether the AI components, including machine learning models and rules engine, were sound in nature and well-built.

The interview questions were designed based on six key areas namely, the logic of making the choice of the offset well, automatic extraction of risks in line with daily reports on drilling, preliminary recommendations of the parameters, according to the AI, the rules-based governance engine, the overall balance between automation and human control, and the greatest risks or unaddressed aspects of the framework. The purpose of the study was explained to all the participants prior to the interview. Participants were volunteers and anonymous. No proprietary information existed on the well or company specific operation statistics of the company that was sought which was well within the ethical considerations in Section 3.2.

Primary Data Survey Questionnaire

In addition to the expert interviews, a structured survey was dispatched to drilling engineers as a method of getting wider feedback on the framework relevance in practice. The questionnaire was developed to investigate their opinions regarding ten key points in the suggested framework using the five-point Likert scale of Strongly Disagree, Strongly Agree. Drilling engineers who are familiar with well program preparation also formed the target population.

This survey was administered online using Google Forms and the target population was to get at least 25 drilling professionals to respond to the survey. The acceptable responses obtained were 22 which is considered reasonable in such conceptual validation study. The responses are analysed and presented in Chapter 4. The fact that both the interview and survey methodology are combined ensures that the framework is not merely tested in terms of the theoretical standards, but also what the engineers that would be using it would want it to be like (Saunders, M., Lewis, P., and Thornhill, A. (2019)).

3.) Data Analysis Techniques

Thematic Analysis (Secondary Data)

The thematic analysis is a relatively trendy method of qualitative research that encompasses recognition, coding and elucidation of sense patterns in a data set through the assistance of the information embodied in the literature review (Braun, V., & Clarke, V. (2006)). The thematic analysis is the most appropriate to the situation because it does not permit the researcher to summarise the individual researches but in the real sense, to develop the underlying themes and the relationship between different pieces of evidence that are intertwined. This is required to create a conceptual framework since the framework must possess the ability to synthesize the data of different disciplines in a model of reasoning and system.

The thematic analysis that is used in the research is based on the six-phase model by Braun, V., and Clarke, V. (2006). The former is the familiarisation of the data by the researcher as he reads and reads the documents he/she selected to have a complete picture of the subject matter of the research. The second one is that of formulation of initial codes that are short names that reflect significant concepts or ideas in the data. The third one is theme-hunting that involves clustering the similar codes into higher-order themes. The fourth step involves a revision and adjustment of the themes so that it is consistent and can be used in the data. The fifth and sixth steps involve defining and writing up of findings respectively. Thematically analysis is conducted under NVivo qualitative data analysis software (figure4) which helps in the rigour and efficiency of thematic analysis (Bazeley, P., and Jackson, K. (2013)).

Qualitative Content Analysis (Expert Interview)

The analysis of the responses of the expert interviews was conducted using qualitative content analysis. All the answers were read and coded based on the critical areas of the interview questions. Where there was an agreement of the professionals or similar issues were raised, that was pooled as a common discovery. In cases where the specialists held special or incongruent views, these were highlighted as important nuances. The results of the interview are presented in Chapter 6.

Descriptive Statistical Analysis (Survey)

Simple descriptive statistics are applied to analyse the survey data. The figures and percentage of respondents who provided each of the options as Strongly Disagree, Strongly Agree etc. are ascertained and summarised against each question. The overall tendency in the answers is used to find out the agreement of the drilling engineers with the idea, that the proposed framework is a solution to the real problems in their current working process. This plan is in line with a common use of survey information in engineering studies where conceptual models are to be verified (Creswell, J. W., and Creswell, J. D. (2018)).

3.3.4 Interview Protocol, NVivo Validation, and Survey Mapping

Interview Protocol

The interview protocol was designed to ensure that expert feedback was directly connected to the proposed drilling program framework rather than general views on artificial intelligence. Each participant was first introduced to the framework diagram and its main components, after which the discussion was guided using structured questions focused on key areas such as offset well selection, extraction of historical risks, parameter recommendations, and the role of safety rules. This approach allowed the researcher to evaluate each part of the framework individually while maintaining a clear link to the overall system. Two groups of experts were involved, including drilling engineers and digital or AI specialists, to capture both operational experience and technical feasibility. All interviews were conducted face-to-face, which enabled clarification during the discussion and helped generate more detailed, experience-based insights. This method ensured that the feedback was practical and directly relevant to the development and refinement of the proposed framework.

NVivo Validation

The interview responses and selected literature were analyzed using thematic analysis supported by NVivo software figure 4. The coding process began by identifying key ideas related to the research objectives, including data challenges, risk identification, model application, and governance. These codes were then reviewed and grouped into broader themes that reflect the structure of the proposed framework. To iimprove the reliability of the analysis, the coding process was repeated and compared across different data sources, including literature and expert feedback. Differences in expert opinions were retained and considered as meaningful findings rather than removed, as they highlight practical challenges and limitations. This approach ensured that the final themes were not only consistent but also grounded in both theory and practice, providing a strong basis for the

framework design (Braun & Clarke, 2006; Bazeley & Jackson, 2013).

Survey Mapping

The survey was developed to support the qualitative findings by collecting feedback from a wider group of drilling engineers. Each survey question was linked to a specific component of the framework, such as challenges in offset selection, difficulty in identifying historical risks, acceptance of AI-generated recommendations, and the importance of safety validation. This structured mapping ensured that the survey results could be directly used to evaluate the relevance and practicality of each framework module. By aligning the survey design with the research objectives and framework structure the responses provided targeted evidence rather than general opinions.This strengthened the overall analysis by confirming whether proposed framework addresses real operational challenges identified by practitioners (Creswell & Creswell, 2018).

C. Comparative Analysis

To further justify the methodology, comparative analysis is added to the research design. Such methodology will be used so as to make comparisons between the traditional, manual approach to the drilling program development and the proposed AI model. By the comparative analysis of these two approaches, the study demonstrates that there are some areas where the conceptual framework would add some value particularly on areas of time efficiency, data consistency and risk reduction.

The comparative analysis is made on the basis of three criteria:

1. **Process Consistency:** When compared to a manual interpretation by individual engineers, examination of the variability of the manual interpretation is a comparison with less heterogeneous output of a rules-based AI selection.
2. **Knowledge Transfer:** Compared to the manual process of locating the offset reports, the automated collection of previous signal history (Amount of mud loss in such a section, what type of BHA).
3. **Auditability:** Measuring the ease of the traditional and proposed digital workflow with respect to the capability of tracking the decisions applied to the planning process to source data.

The logical rationale of the creation of the framework is provided through this comparative strategy. It assists in ensuring that the proposed solution does not transform into an easy technological supplement to the existing manual procedures that are cited to be ineffective. The outcomes of this comparison are the background of the validation stage in Chapter 5.

V. ANALYSIS

This chapter interprets the evidence gathered for the study and shows how that evidence answers the two research questions. The conceptual findings in Sections 4.2 to 4.4 come mainly from Chapter 2, especially the themes on digital well planning, hierarchical offset selection, machine learning for drilling parameterand rules-based governance. The survey findings in Section 4.5 come from the primary data process described in Chapter3 where the questionnaire design, respondents, and descriptive statistical approach were explained. Chapter 4 therefore links the theoretical foundation in Chapter 2 the research process in Chapter 3, and the framework developed in Chapter 5.The analysis indicates that the proposed framework is not centred on automation alone. It addresses a practical planning difficulty: drilling engineers often spend valuable time searching for offset wells, reviewing old reports, and trying to recover lessons that already exist but are not organised for quick reuse. The findings point to three needs: better offset selection logic, better data structure, and stronger governance over AI-generated outputs. In addition, qualitative insights from the literature and survey responses were supported through thematic structuring using NVivo, which assisted in identifying recurriing patterns related to offset selection, data challenges, and governance requirements.

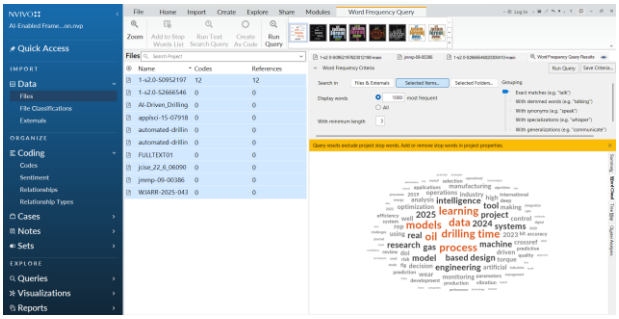


Figure 4 Word Cloud

A. The Industrial Need and Conceptual Foundation

The analysis of the research codes revealed that industrial demand was too much on the need to move towards the "Smart Oilfield." The statistics show that much is already being wasted within the industry by wasting invisible time, lack of data of experts. According to Williams (2022), the inability to identify high-quality labeled datasets is also the

initial impediment to improving the application of AI to engineering design as discussed in chapter 2. In addition, according to the research by Adjei (2025) and Sircar (2021), discussed in chapter 2. Manual planning is no longer sufficient in the depth of the contemporary reservoirs. According to the statistics, even such giant players in the market as Shell and Chevron have already achieved efficiency gains of up to 130 percent through the implementation of AI-enhanced optimization models. The most interesting outcome of this analysis is also the role of the professional that is also changing, as defined by Koroteev (2021) as the transformation of the engineer. The data refer to the fact that AI does not replace the engineer but rather only alters his or her work to the extent where he/she can now be an AI manager and be capable of focusing on making the high-level decisions, rather than data searching with a great deal of effort chapter 5.

B. Data Integration and Hierarchical Connectedness

To provide answers to the first research question on how offset wells are systematically utilized, data identification and processing had been analyzed. This information verifies that scientifically most reasonable procedure to select hierarchy offset is the one, which is unable to discriminate the similarity of reservoirs and subsequently the provisions of spatial radius. It is essential to mention that to execute an AI structure, Williams (2022) indicates that the historical report must be transformed into machine-readable forms and standard forms of metadata as discussed in chapter 2. Without this transformation, AI systems cannot effectively analyze historical failure of operations (stuck pipes or circulation losses). The other notable finding of the analysis was that there was an urgent need of what is referred to as a cyber-physical pathway. This programming fibre links the planning document with the actual real-life rig protocols and hardware interfaces. Khan (2025) explains that the adherence to technical standards like the WITSML and OPC-UA would be a viable solution in ensuring that the parameters that were employed during the planning phase would be compatible with the physical rig equipment at the wellsite.

C. Artificial Intelligence Regulations of Governance and Modeling

To respond to the second research question that is regarding consistency and quality the analysis took into account interaction between machine learning and engineering rules. The facts point to the fact that the combination strategy is superior to the application of single algorithms. Mohammadinia (2025) states that Gradient Boosted Trees are highly fruitful in the role of learning models to predict the speed of penetration, whereas on the one hand, reinforcement learning and Bayesian Optimization is required to keep these parameters within the harmless work ranges as discussed in chapter 2. A critical implication is that AI cannot be implemented to manage safety related operations in the oil and gas sector by itself. It demonstrates that a rules-based engine of governance must be considered the last gatekeeper, as demonstrated by Mertiri (2025) and Gao and Sun (2023). This engine applies corporate safety standards and regulatory limits that provide the traceability and structural completeness that a manual program may not provide. Besides, the threat of algorithmic bias and conceptual drift is also detected in the process of analyzing Povooas (2025), hence the need to monitor the model at all times to maintain the model structure as both ethical and technically sound over an extended period.

Table 1Comparative synthesis of literature findings and implications for the proposed framework

Theme	**Evidence drawn from Chapter 2**	**Main limitation in prior studies**	**Implication for Chapter 5 framework**
Digital well planning	The review linked AI adoption to efficiency, reduced NPT, and better use of historical knowledge.	Most studies describe digital progress broadly but do not convert it into a full planning workflow.	The framework should be presented as a practical planning system, not only a technical idea.
Offset selection	Relevant studies supported reservoir-based relevance before simple proximity.	Manual selection remains subjective and many studies stop at analogue identification.	Offset wells should be ranked by reservoir first and distance second before downstream analysis begins.
Prediction and optimisation	The literature justified the use of predictive ML together with optimisation under constraints.	Most models address one task at a time and are not connected to the final planning document.	AI modules should support parameter estimation and optimisation, but as part of one linked workflow.
Rules-based governance	Several studies argued that safety-criticaldrilling decisions require explicit engineering rules.	Governance is often discussed conceptually and not embedded into model outputs.	A rules engine should checkrecommendations against standards before engineer review.

D. Comparative Synthesis:

The comparative logic outlined in Chapter 3 explains why this matters in practice. Manual planning can still produce strong results, but quality depends heavily on who performs the work, how much time is available, and whether historical lessons are located and interpreted consistently. A governed

AI-supported workflow does not remove engineering responsibility, but it reduces variation in selection, extraction, and checking.

E. Survey Results

A survey was developed to supplement the secondary literature analysis by interviewing drilling engineers on their perception of the problem of the existing manual planning process and whether the proposed AI framework would solve the problems. The questionnaire had ten items with a five-point Likert scale. Professionals in drilling were surveyed and gave 22 answers. Below, the findings will be summarized into three themes, namely, the current issues in manual planning, the practicality of particular features of framework, and the general worth of the framework.

1) Existing Problems with Manual Planning

Question 1: Minutes spent in searching data manually

The respondents were required to answer questions on whether they take a lot of time manually searching and examining historical offset well data. The results showed that 8 respondents Agreed and 8 Strongly Agreed bringing the total agreement to 72.7 percent. Strongly Disagreed and Disagreed only 4 and 1 respectively. This is in support of a direct and consistent literature conclusion: data hunting becomes a factual load and a time-consuming cost to the drilling engineers.

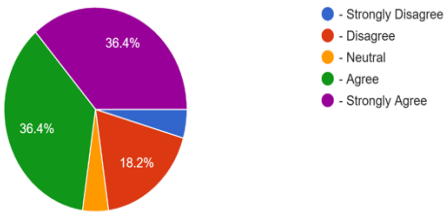


Figure 5 Searching data manually

Question 2: The subjectivity of the selection of offset wells.

Responding to the question of whether not or whether or not the selection of offset wells is subjective and may vary among the engineers when selecting the same new well, 11 respondents Agreed and 4 Strongly Agreed and gives an agreement rate of 68.2. There were only 6 Disagreed and 1 was Neutral. This is to advance the argument that there is inconsistency in the current manual process, and that there exists a need to have some form of structured selection hierarchy. This is an observation that is consistent with the problems raised in the literature where manual planning is extremely reliant on individual experience and judgment.

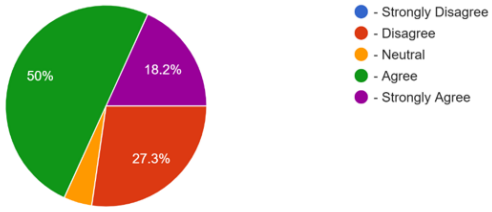


Figure 6 s=Selection of offset wells

Question 3: Difficult to get historical risks

The respondents were questioned on whether it is difficult to systematically draw out major risks like stuck pipe or circulation losses that were reported in the past in a new plan. This was the highest rate of consensus of all questions to be challenged: 11 Agreed and 7 Strongly Agreed with the overall agreement rate being 81.8%. The number of Disagreed and Neutral was only 3 and 1 respectively. This result directly confirms one of the main issues that this study will address, and coincides with the literature finding that the risks that have been identified are rediscovered rather than avoided during the planning stage.

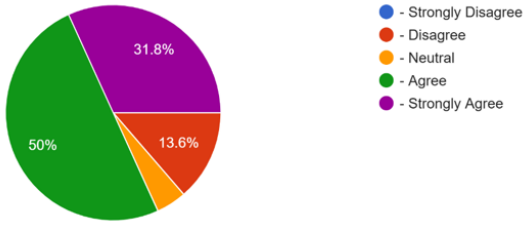


Figure 7 Historical Risks

2) Applicability of the Specific Frameworks Features

Question 5: Automatic risk flagging on the daily drilling reports

The participants were asked to answer questions regarding usefulness of the automatic flagging of historical risks involved in daily report work at various depths of a new plan. 10 Agree and 8 Strongly agree, with 81.8% total agreement. 4 Discords and no one was a Neutral type i.e. there was no gray zone between the two extremes i.e. respondents either agreed or disagreed on this feature. This is a strong result that supports the value of the NLP founded risk extraction component of the framework.

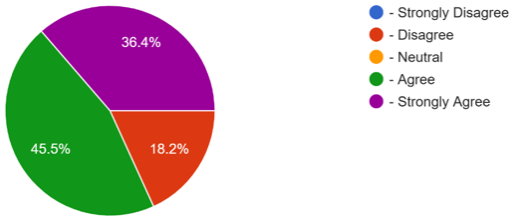


Figure 8 Automatic Risk Flagging

Question 8 Liberating engineers to concentrate on judgment and not data

Since the respondents were inquired about whether an integrated workflow would allow them to spend more time on engineering judgement and less on data hunting, 9 Strongly Agreed and 8 Agreed, which amounts to 77.3% of the respondents agree. 4 dissented and 1 was Neutral. The observation would align with the thesis statement that AI should be used to assist, but not replace, the engineer and represents the overall trend in the industry as described in the literature of human-AI planning, and not full automation.

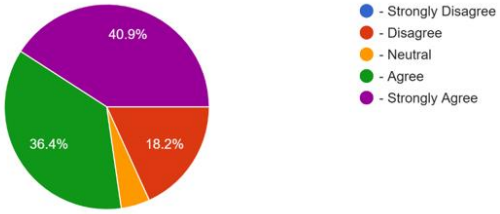


Figure 9 Liberating engineers to concentrate on judgment and not data

Table 3 Summary of Survey Findings

Q#	Topic	% Agree or Strongly Agree
1	Time spent on manual data hunting	72.7%
2	Subjectivity in offset selection	68.2%
3	Difficulty extracting historical risks	81.8%
4	Hierarchical offset ranking tool	77.3%
5	Auto-flagging historical risks	81.8%
6	AI-generated parameter recommendations	81.8%
7	Auto-verification against safety rules	81.8%
8	Focus shift to engineering judgment	77.3%
9	Traceable outputs for audit	77.3%
10	Overall improvement in planning quality	81.8%

VI. PROPOSAL FOR AN AI-ENABLED FRAMEWORK

Even though, the proposed framework is hypothetically developed to be inclusive of all the twelve technical areas required by a drilling program- this study narrows down to

three key areas of high impact. The specific modules are selected to provide the study with a direction and ensure that the research approach becomes narrowed down and can be measured and qualified in line with the available manual provisions.

The area of central focus in this study is:

- Automated Offset Well Selection**:** Replacing manual searches (that are subjective) with a hierarchical and multi-criteria similarity analysis.
- NPT and Risk Extraction of Daily Operations Reports: Replacing active planning (reading past programs) with a proactive step, which is reading daily operational reports and learning what past problems (e.g. stuck pipes, losses) actually happen that will not be reflected by the planned programs.
- Optimized Casing and Mud Program Design: Bayesian Optimization and Machine Learning to the two most significant technical elements of the project to ensure that they are coordinated and safe, which are now highly diverse between the engineer-to-engineer.

These three areas will enable the study to quantify the benefits on time savings, consistency on the basis of the data and the risk awareness compared to the more traditional manual-based practices with the highest degree of accuracy.

A. Framework Development

AI Back-End Multi-stage Processing Workflow

By the framework architecture of the AI back-end, it is processed on a multi-stage processing pipeline with structure. One stage is fed on some input data and the output data feeds on other stages and downstream planning modules. The

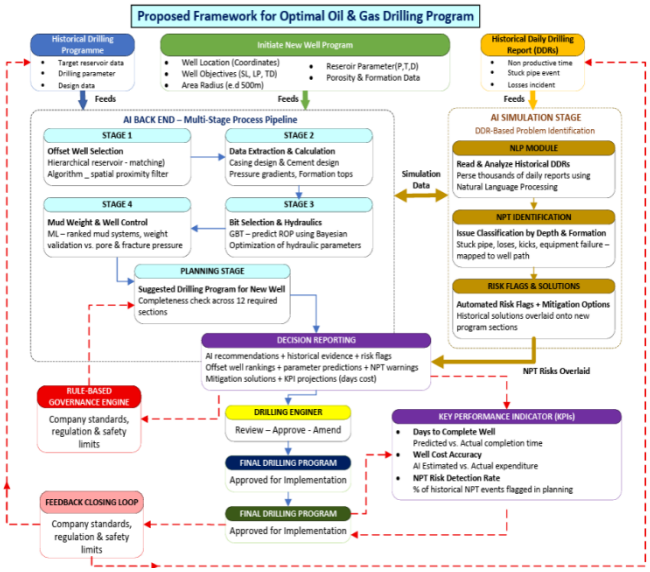


Figure 1 Framework

workflow is as follows:

Input Stage: Data Ingestion:

The engineer feeds the target well position figures (latitude, longitude), significant reservoir factors like porosity (P), temperature (T) and depth (D) and a search radius about the target (e.g. 500 metres). The AI back-end is triggered by the inputs that are entered into the system.

Stage 1: Geological Algorithm and Offset Well Selection:

The algorithm is hierarchical reservoir-matching algorithm that initially consults wells within the identical geological structures and reservoir characteristics and the spatial radius filter ranks the offset wells based on their distance. This will form a geologically and operationally most relevant standardised list of the offset wells (AWE -Automated Well Evaluation Preparation).

Stage 2: Data Extraction and Calculation of parameter:

Depending on the offset wells, which will be picked, the system will retrieve the past drilling parameters that include casing logs, completion logs, formation tops and performance indicators of the undertakings. Systematic calculations that give the most important calculations (count of returns, casing setting depths, and pressure gradients) are made on the basis of extracted data.Stage Stage 3: Mud Weight Inquiry and Well Control Test:

The framework questions the historical mud programmes of the similar offset wells, retrieves the mud systems, which were most performing, and cross verifies mud weight propositions against pore pressure forecasts and fracture gradient forecasts to ensure that there are adequate well control margins.

Stage 4: Comparison and Check of Completeness and Planning Sheet Generation:

All the parameters that have been extracted and optimised are summarised into one planning sheet. It carries out a completeness check of the twelve required sections of a standard drilling programme and where the information is insufficient or where there is need to apply engineering judgment is denoted. The result is a Decision Report that internalizes the AI suggestions and historical evidence of the suggestions.

B. Problem Implementation and Simulation Step DDR-Based

Parallel to the multi-stage pipeline, the framework also executes another special Daily Drilling Report (DDR) analysis module. This module is linked with the historical database of DDRs of the selected offset wells and applies the algorithms of Natural Language Processing (NLP) to the identified issues that had been registered, the NPT events, and the operational problems on specific depths and formations.

The problems mentioned are classified and ranked by type (stuck pipe, losses, kicks, equipment failure), depth interval, and formation and overlaid onto the new-created well plan as automated risk flags. This is the Simulation Stage of the framework because the AI models the potential problem areas, by superimposing the data of the past occurrences over the intended well path.

Feedback Loop Mechanism

The feedback loop is extremely important element of the framework as shown in the architectural design. When drilling programe is implemented and actual drilling of the well is in progress, the real data of operation: the final Daily Drilling Reports, the actual costs and the time it took to complete, the number of the NPT events incurred is inputted into the database of the historical data. This feedback enhances the data to be employed in the future in an offset well query so that the AI models can evolve to the current

operational outcomes and continue to enhance the accurateness of the predictions and recommendations to the future wells of the identical field or formation.

C. 5.3 Offset Well Selection

Manual Practice: The engineers search through databases of historical wells records, manually, finding the offset well in terms of personal knowledge, position and available records. It is also time-consuming and fails to provide similar results among engineers hence reference wells selection to the same planned site is not uniform.

AI Conversion: The structure will entail automated and hierarchical well offset well selection module where queries of initially wells in the same reservoir and then followed by a determined spatial range is sought and interrogated in case the wells are not located in the determined spatial range. It will also ensure that the individual requirement is never adopted to determine the most geologically and operationally viable wells (Feng and Han, 2015). The engineer puts in the target well coordinates and a search radius (e.g. 500 metres) and the AI back-end Stage 1 algorithm delivers a ranked list of offset wells filtered by reservoir match and spatial proximity which is directly inputted to all downstream planning modules.

D. Casing Design Parameters

Manual Practice: Engineers can determine the casing setting depths, size, grade, and design loads by manual calculation to burst, collapse, and tension basing their decision on the calculated profiles of the expected pressure and the anticipated downhole condition. Various safety factors or design assumptions can be used by every engineer.

AI Conversion: The model input is the parameters of casing design of successful offset wells under the same pressure and geological setting, and it applies Bayesian Optimisation to determine optimal casing designs within safe operating space provided by company specifications and regulatory guidelines. The rules engine uses minimum design requirements and complete structure checks to provide recommendations to the engineer (Teixeira and Secchi, 2019).

E. Drilling Mud Program

Manual Practice: The engineer determines the requirements made by the mud weight, the fluid that will be employed, specifications of the treatment and even compatibility with the equipment by examining the formation properties, the anticipated pressures, and having a previous experience with similar formations. The choice is normally determined by the personal preference and a prescribed understanding of the application of particular mud systems.

AI Conversion: The systematically accessed mud program data of the offset wells, which drilled in the same formations, which mud system had the highest success according to the hole stability and effectiveness of drilling, and suggests the most desirable mud properties, according to the machine learning models. The governing rules are such that rules of the recommended mud weight ensure that there is an adequate well control margin without fractured formation strength (Al-Rubaii et al., 2023). The module relates to Stage 3 of the AI back-end pipeline where the queries of mud weight are compared with the pressure prediction queries of Stage 2.

F. 5.6 Bit Selection and Hydraulics

- Manual Practice: The choice of the bit is made according to the performance record of the neighboring wells, formation drillability, mud system compatibility and directional requirements. The computation of hydraulics of annular velocity, Bit horse power hydraulic and jet impact is performed manually.
- AI Conversion: The model is anchored on the historical record of bit performance on offset wells and is used to predict the type of bit most suitable and each section of the hole hydraulics. The Rate of Penetration of different combinations of bits and parameters is estimated with the help of Gradient Boosted Trees and the maximisation of the number of parameters leading to the safe drilling operation is determined through the application of Bayesian Optimisation (Mohammadinia et al., 2025).

G. NPT Prevention and Risk Identification

- Manual Practice: At present engineers primarily examine the completed " Drilling Programs of past wells to identify risks. But these programs do not portray what happened but what should have happened. In case an issue such as a clogged pipe arose, it could be buried in a mountain of thousands of pages of handwritten or typed Daily Drilling Reports (DDRs). This is Active Planning--it adheres to a plan but it usually lacks real life challenges that are encountered by the former crew.
- AI Conversion: Natural Language Processing (NLP) is used to convert the framework to Proactive

Planning. NLP is an intelligent technology based on AI that is capable of reading and comprehending thousands of Daily Drilling Reports within several seconds. The AI does not examine the old plan but scans the real-life comments of past engineers on a daily basis to determine where and why Non-Productive Time (NPT) occurred.

When these lessons learned are discovered based on real world information on a daily basis, the system automatically recognizes threats like specific depths in which the company incurred a loss and immediately inserts them into the new program. This would ensure that the engineer would be made aware of the actual facts of history and not optimal plans and the new well would be much safer (Salem et al., 2022). This module is based on the DDR historical database and supplies its results to the Simulation Stage where the NPT patterns identified are overlaid on the planned well path as depth-specific and formation-specific risk indicators.

H. 5.8 Human controls and Regulatory Governance

The most critical thing about the proposed framework is that all outputs obtained by an AI are processed by a rules-based governance engine, which is presented to the drilling engineer. Such an engine will also make sure that the company drilling standards, regulatory requirements, and safety constraints are followed. It also does structural completeness to determine that all the required areas of drilling program are addressed. The framework does not give final rulings and does not pursue any business. All the outputs should have their recommendations revised, approved or modified by qualified engineers who shall be fully accountable of the professional responsibility in the final drilling program. This is the extent of governance by the proposed framework, as compared to the full autonomy systems, and ensures that the benefits of AI are realized in a regulated and auditable engineering scenario (Arinze et al., 2024).

I. 5.9 Problem Solution Workflow and Decision Report

The framework leads to a Decision Report that is the compilation of all AI-generated recommendations, historical data, and risk flags into a report that is now ready to review by the drilling engineer. The Decision Report will include:

- Ranking of offset well locations on justification statistics (reservoir match score, distance)
- Past confidence intervals of the expected drilling variables.
- NPT risks that were identified were used on the depths and formation of the planned well.
- The recommended solutions to the mentioned problems are pegged on the successful mitigation procedures articulated in the historical DDRs of the same offset wells.
- KPI projections like days to complete projections, cost forecasts which were compared against offset wells actuals.

The engineer will discuss the Decision Report, make a professional decision and approve / amend the recommendations and sign the final drilling programme. Any alterations as may be made by the engineer are documented to allow the engineer to trace them and are sent back into the system to feed the later suggestions.

VII. DISCUSSION AND THEORETICAL IMPLICATIONS

A. From Fragmented AI Point Solutions to Integrated Rule-Based Governance

This paper shows that there is a need to be developed in terms of the application of digital technology in the drilling sector. Historical artificial intelligence and machine learning in drilling have been considered point solutions, meaning, one specific task, i.e. the rate of penetration, or when a pipe is stuck. Even though these tools are very helpful, they are at times isolated and the engineers have to fill the gap between the various results manually. This conceptual framework created in this study is dynamic and thus changes this by providing an umbrella design due to which these individual tools are co-ordinated. The validation phase discussion with the industry authorities confirmed that the true value of AI should not be limited to its predictive capacity, but should have the capacity to organize and present the past in such a format which can be acted upon by an engineer in the planning capacity in an immediate manner.

One of the most important lessons that were made during this work is the importance of the hierarchical offset selection logic. The existing analog planning theories suggest that proximity is the major consideration to be made when choosing offset wells. Yet this framework was developed and constructed and indicated that the geological relevance, which is the connectivity of reservoirs, must be viewed as the

priority of what simple geographic distance can provide. This provides a twist into the traditional neighbor proximity strategy. Paying attention to the wells that lead to the same reservoir, the AI framework will be capable of making the parameters proposed by the AI framework, such as the mud weights and depths of the casing reliant on the most relevant physical conditions, rather than on the closest values available. It was found that this hierarchy reduces the noise of irrelevant data that is a common problem in mature fields where hundreds of wells can be located within a small area.

Besides, a rules-based engine of governance also addresses a key theoretical gap in AI-based studies. Majority of machine learning models have been criticized as black boxes where the logic behind a prediction can be not seen. A black-box solution is not accepted by the top management in such a high-stakes situation as the oil and gas drilling, when one misstep can lead to the environmental disaster or the loss of money on a massive scale. This is replaced by the structure discussed herein, which has an open rules engine as the final filter. This causes any AI-generated suggestion to be checked against a predetermined safety standard and regulation. It has a two-layered model that is rooted in machine learning to optimize and rules to be safe; it is new and trustworthy. It is also aligned with the idea of human-in-the-loop AI in which the technology will act as a complement to the expert, rather than replace him or her.

To further support the discussion quantified estimation of the expected performance improvements of the proposed framework is presented. Based on survey findings and expert feedback, the framework is expected to reduce drilling program preparation time by approximately 30% to 50%. This is mainly due to automation of offset well selection and structured extraction of historical data which addresses inefficiencies commonly associated with manual data searching and report review (Rahman et al., 2023; Petrofac, 2024). In addition the framework is expected to improve planning completeness and consistency by approximately 20% to 30% as the rules based governance layer ensures that all required elements of the drilling program are systematically validated rather than depending on individual experience (Gao & Sun, 2023; Arinze et al., 2024).

From a risk management perspective the integration of automated extraction of historical events from Daily Drilling Reports is expected to increase early stage risk identification by approximately 40% to 60%, particularly for non productive time events such as stuck pipe and losses. This is consistent with previous findings that highlight the lack of systematic risk transfer in conventional planning workflows (Amani et al., 2024; Salem et al., 2022). In a simulated planning context, this would lead to a higher probability of identifying known failure zones before execution, reducing the likelihood of repeated operational issues. Although these values are based on conceptual estimation rather than field implementation, they provide a realistic indication of the framework's expected impact on efficiency, quality, and risk awareness.

To further demonstrate the alignment between the research problem, objectives, and research questions Table4 provides a structured mapping of how each element has been addressed throughout the thesis.

B. 6.3 Reflection on Research Questions

The initial research question was how the use of the offset wells in a typical reservoir or a certain radius could be concerned in the planning parameter and risk determination in a systematic manner. The paper found that the most appropriate of this is by a hierarchical selection process. The framework is able to automatically create casing points, mud weight envelopes and historical non-productive time events first of all, by filtering on the basis of the reservoir and secondly by filtering on the basis of spatial proximity. The study has shown that the omission of lessons present in the manual planning is avoided by the planned extraction. The AI can also make all of the historical events visible and offer them as a risk annotation, which would not happen in case an engineer would not look at a report that was written five years ago. It is an answer to the first question because it provides an automated logic of transferring data.

The second research question was what was the degree to which a structured framework is more superior to manual preparation in terms of consistency, traceability and quality. The comparative analysis and the analysis of the experts made it possible to determine that the improvement is significant. No one is also objective as manual planning because the same data on offsets may be interpreted differently by the various engineers depending on their experience. The homogeneityof the reasoning is introduced by the AI system. As the system uses the same rules and algorithms on all wells, the programs obtained in the process would be the same in all parts of the organization. There is also a high amount of traceability. The tendency in manual is that an individual may not be aware as to why a certain amount of weight of mud was chosen. In the proposed scheme, all the parameters are linked to either a specific offset well or a rule to be followed, and therefore there would

be an evident audit trail. This determines the fact that the framework to a great extent enhances the standard and dependability of the drilling program.

C. Integration of Primary Data Findings

1) Expert Interview Findings

The structured interview conducted with a drilling expert gave first hand practical feed back on the framework proposed. The general evaluation of the expert has proven to be positive and there is proof that the framework was well-organized and comprehensive. However, some of the key observations were made that have shaped the final version of the framework.

The expert accepted that hierarchical offset selection strategy is more reliable and justifiable since by this method, first hierarchically by reservoir, and then hierarchically by spatial proximity, removes the more tedious and arbitrary existing selection techniques, which are manual. This goes to the point of directly confirming the Research Question 1 that needed to determine the way the utilization of offset wells can be conducted in a systematic manner in order to obtain planning parameters and risks. The hierarchical organization was described to reduce bias, make the logic of selection transparent or easier to be explained during the peer review or audit. This is in line with the literature argument that relevance of reservoirs be accorded a higher priority than that of the geographic distance when selecting offset wells. (2023).

Answering the question whether it is possible to scan daily drilling reports using Natural Language Processing and determine whether some issues have been detected before at some depths, the professional stated that such a measure would go a long way in eliminating the repetition of problems. The history of the many issues is already captured in reports and they are rarely reconsidered systematically in the planning process. This is because when these risks are seen in the planning process and not found in the implementation, institutional learning and risk mitigation is improved. This corroborates the findings of the literature according to which the systematic transfer of risk based on the information of offset is among the most significant missing links in the traditional planning processes (Amani, M., et al. (2024).

On AI-generated initial parameter recommendations, the expert accepted that there is high value in offset performance successful to create initial recommendations on casing points, mud weights, and bit selection. The expert, however, noted that such outputs can be regarded as preliminary assumptions, as opposed to the final engineering decisions particularly in cases where there is geomechanical complexities. This substantiates the selection of the framework design to make AI outputs present itself as recommendations that can be discussed by the engineer rather than as solutions. (2023).

The AI expert confirmed that the automatic checking against the company standards and safety envelope of AI findings would significantly increase the consistency and reduce the chances of the presence of loopholes in compliance that can be achieved due to time constraints. This takes the work of engineers out of the normal compliance check as far as higher level engineering judgment where human expertise can add the most value. This is in line with the literature that proves the existence of a required rules engine to support safety constraints and provide traceability in AI-assisted planning (Al-Dushaishi, M. F., and Al-Sofi, H.). (2025).

Data quality and contextual interpretation was also the most significant risk of the framework as determined by the expert. The language of Daily Drilling Report might be incoherent, incomplete, or vague. Inability to explain clearly the confidence levels, assumptions, and gaps in the data in the system may lead to excessive trust in the system outputs by engineers. This feedback is now socialized in the framework design that now obviously requires transparency indicators and AI-generated advice.

Some of the structural recommendations that the specialist gave included restructuring the arrangement of the elements of the framework. The specialist suggested the following correct flow to be: AI Recommendations -Rule-Based Governance -Engineer Review - Final Program. This ensures that the engine of governance is used as a filter to the engineer looking at the output and not as a parallel or secondary process. This conducted in a systematic manner in order to obtain planning parameters and risks. The hierarchical organization was described to reduce bias, make the logic of selection transparent or easier to be explained during the peer review or audit. This is in line with the literature argument that relevance of reservoirs be accorded a higher priority than that of the geographic distance when selecting offset wells. (2023).

Answering the question whether it is possible to scan daily drilling reports using Natural Language Processing and determine whether some issues have been detected before at some depths, the professional stated that such a measure would go a long way in eliminating the repetition of problems. The history of the many issues is already captured

in reports and they are rarely reconsidered systematically in the planning process. This is because when these risks are seen in the planning process and not found in the implementation, institutional learning and risk mitigation is improved. This corroborates the findings of the literature according to which the systematic transfer of risk based on the information of offset is among the most significant missing links in the traditional planning processes (Amani, M., et al. (2024).

On AI-generated initial parameter recommendations, the expert accepted that there is high value in offset performance successful to create initial recommendations on casing points, mud weights, and bit selection. The expert, however, noted that such outputs can be regarded as preliminary assumptions, as opposed to the final engineering decisions particularly in cases where there is geomechanical complexities. This substantiates the selection of the framework design to make AI outputs present itself as recommendations that can be discussed by the engineer rather than as solutions. (2023).

The AI expert confirmed that the automatic checking against the company standards and safety envelope of AI findings would significantly increase the consistency and reduce the chances of the presence of loopholes in compliance that can be achieved due to time constraints. This takes the work of engineers out of the normal compliance check as far as higher level engineering judgment where human expertise can add the most value. This is in line with the literature that proves the existence of a required rules engine to support safety constraints and provide traceability in AI-assisted planning (Al-Dushaishi, M. F., and Al-Sofi, H.). (2025).

Data quality and contextual interpretation was also the most significant risk of the framework as determined by the expert. The language of Daily Drilling Report might be incoherent, incomplete, or vague. Inability to explain clearly the confidence levels, assumptions, and gaps in the data in the system may lead to excessive trust in the system outputs by engineers. This feedback is now socialized in the framework design that now obviously requires transparency indicators and AI-generated advice.

Some of the structural recommendations that the specialist gave included restructuring the arrangement of the elements of the framework. The specialist suggested the following correct flow to be: AI Recommendations -Rule-Based Governance -Engineer Review - Final Program. This ensures that the engine of governance is used as a filter to the engineer looking at the output and not as a parallel or secondary process. This repositioning has found its reflection in the new framework diagram in Chapter 5.

2) 6.4.2 Survey Results and Implications

The general quantitative confirmation of the framework is the survey results of 22 drilling engineers. The overall mean agreement was approximately 78 in all ten questions and no question had lower average than 68. These results provide a substantial support that the framework will discuss real operation loopholes identified by practising engineers.

The results confirm the validity of the data hunting problem in that 72.7 percent of the participants have affirmed that they devote a lot of time in the manual search of past offset data as discussed in chapter 4. This goes hand in hand with the fact that in the literature it is observed that drilling engineers spend disproportionate time of their planning time in finding offset wells and reading old reportages instead of conducting engineering analysis. (2023)). The fact is that, the subjectivity of the offset option was confirmed by only 68.2 percent of the respondents, which is the lowest of all questions as shown in chapter 4. It is an implication that engineers are mostly aware of the inconsistency problem, but can think that their personal practice is consistent enough. This is among the potential adoption barriers: engineers who think that they do a good and efficient job will be less motivated to have different attitudes.

The maximum consensus of 81.8% was reached in four features, i.e. automatic extraction of historical risks, parameter recommendation defined by AI, automatic checking of safety rules, and the idea of the overall framework. The fact that most of the respondents agreed with the safety rule checking option almost unanimously (81.8 per cent agree, zero neutral) is the evidence that this tier of governance is not only theoretically correct but is sought after by the practitioners in practice. This confirms the literature observation that a rules engine offers the structural completeness and traceability that is not available to manual planning at the moment. (2023).

VII. IMPLICATIONS FOR PRACTICE AND FUTURE RESEARCH

There is a need to have a good appreciation of the weaknesses and the strength of any solution proposed to re-align the oil and gas industry between the abstract theory and the implementation in the real world. This chapter will give the practical implication of the study on energy companies and drilling engineers based on the findings of the literature

review, the professional interview and the survey of the drilling experts. It also outlines the limitations of the study that was conducted and gives suggestions on how the limitations can be overcome. Lastly, the chapter presents the possible future research directions which may involve the creation of a functional software prototype and the incorporation of real-time drilling data into the planning process.

A. 7.2 Implications for Practice

The results of this study have direct and obvious implications on the drilling teams and digital transformation leaders in oil and gas organisations. The hypothesis that there is a high level of practical demand among the engineers who would be the users of the proposed structured AI-assisted planning tool is proven by the results of the survey which showed that on average 78 percent of the interviewed believed that the proposed framework features would be valuable to their work. (2023)). The highest scores of the agreement were achieved in automatic risk extraction of daily drilling reports, automatic recommendation of parameters by AI and automatic checking of safety rules. This informs organisations that it is these three features that need to be given priority in every implementation roadmap.

The expert interview affirmed this by confirming that the greatest real-life benefit of the framework that would be realized fastest would be the removal of time spent on hunting data to enable the engineers to redirect their efforts to a more advanced level of technical analysis and judgment. The expert also noted that the traceable audit trail that the framework would leave behind linking all the recommendations to their source data, ML output, and rule of governance would make programs far simpler to audit and justify in the peer review or regulatory audit. This traceability aspect is adequate in organisations where the circumstances are highly regulated to signify a profound improvement in the current manual documentation procedures (Al-Dushaishi and Al-Sofi, 2025).

The framework can also offer an efficient model of uniformity of quality of planning between locations and teams to digital transformation leaders. This is because the same selection rules and algorithms will apply to all the wells hence the quality of output will not depend on the level of experience of the specific engineer who will be sent to work on the assignment. It is particularly applicable in those organisations, which have many fields/region coverage and cannot easily be consistent in planning through manual means (Petrofac, 2024).

B. 7.3 Limitations and Recommendations

Although the proposed framework has its strong sides, a number of limitations should be mentioned. To begin with, the framework is theoretical and methodological nowadays. It is a description of the logic of the working process, data interactions, and model roles, not working software code or a user interface. This means that its practical implementation, when time limits and incomplete information are involved, and complex geological conditions, are not yet practical.

The framework assumes that historical well files are electronic format, reasonably complete and in a regular format. Practically, the older well records might only be in the form of manual reports or of low quality scans, which cannot be processed by the AI systems without a lot of pre-processing. This was the greatest threat to the success of the framework that the expert interview had found because any system not equipped with the ability to communicate with the engineer the degree of confidence, gaps and assumptions in data is one that is at risk of over-reliance on the outcomes of a system that is driven by incomplete or inconsistent source data partially.

The third limitation, which is indirectly stated by the survey results, is that there may be resistance towards adoption. Although 68.2 percent of the surveyed people consented that the procedure of an offset selection is subjective and inconsistent (the lowest point of all questions), it is noteworthy that some percentage of engineers consider that their current individual practice is adequate. The organisations implementing this framework should not anticipate that all people will buy-in instantly.

To address these weaknesses, organisations have been encouraged to undergo a data purification effort first before any AI implementation. Past reports are to be digitised and standardised with the help of natural language processing software so that the AI has a solid and stable knowledge base to operate on. The introduction of the framework should also take place in stages and not as a wholesale replacement of the current planning processes. In the first step, one can suggest offset wells and mark historical risks through the system and provide the engineers with time to gain some familiarity and trust before the automated parameter recommendation layer is introduced. Training is also crucial: engineers should comprehend the reasoning of the AI proposals to be able to observe the system efficiently and take action in case the situation of geological or operational factors is too far to the right of the historical data (Gao and Sun, 2023).

C. 7.4 Future Work

The next most important thing that this research must undertake is the creation of a fully operational software prototype. An operational digital tool, which would be founded on the conceptual model, would allow conducting a mass back-testing of thousands of historical wells, feeding the machine learning algorithms with real performance data, and estimating the real decrease in the time of planning and the number of failures. This would also allow testing the framework with the specified problems of data quality as one of the limitations which would be empirically tested as to how effectively the NLP-based risk extraction component can process inconsistent or incomplete legacy reports.

The further study will need to take into account the application of more advanced examples of AI such as the Reinforcement Learning that would allow the system to learn as the outcomes of the programs it helps to produce. This would eventually result in a self-refining planning mechanism that would be more accurate and reliable as every well was drilled. At present, this type of feedback loop is not present in any of the point solutions known in the literature review (Allawi et al., 2025).

The second crucial aspect that may be enhanced in further work is the implementation of real-time drilling data in the planning organization. The current system is geared towards the pre-spud planning process. However, a dynamic system that ensures the drilling program is updated as the drilling is being done would be an enormous jump. As an example, in the case of a well hitting a pressure zone above the model, the AI could automatically recalculate the data on the appropriate offset and propose a new mud weight or casing depth in real-time, therefore, reducing response time and increasing well control. Finally, more research could look into how the framework could be extended to incorporate the objectives of environmental and social governance and make AI achieve the maximum optimization of the minimization of a carbon footprint and regulatory adherence to the traditional engineering principles of cost and speed.

VIII. CONCLUSION

The purpose of this thesis was to make and test a theoretical AI-based system to plan the drilling programs in the oil and gas industry. The motivation to carry out the research was that conventional process of well planning is very manual and labour intensive and involves a lot of dependence on experience of individual engineers. Such a process is usually dependent on human memory and manual search of historical information, and therefore can result in an inconsistent plan, the loss of important historical risks, and repetitive operational failures. This thesis explored the latest trends within the area of artificial intelligence, surveyed how the industry was already operating, and collected the first-hand evidence by conducting professional interviews and a systematic survey, which led to the systematic and evidence-based resolution of these problems.

The framework as elaborated in the study is related with three basic innovations in the field of well delivery. Firstly, it is associated with a hierarchical principle of choosing the offset wells in which the geological factors, i.e., the connectivity of reservoirs, is valued more than the geographic distance (Gao and Sun, 2023). This ensures that the information that is used to make decisions on new wells reflect the actual picture of what is happening in the subsurface environment and not the nearest one available. Second, the framework is a hybrid of predictive and optimisation machine learning models to determine the parameters Gradient Boosted Trees and optimisation of the values to the safest and most efficient ones under constrained limits (Mohammadinia et al., 2025; Teixeira and Secchi, 2019). The integration will bring the system to a stage that is not only data retrieval but more of the optimal parameter envelope that is grounded on performance and safety. Third, a business engine of governance offers the required control to make sure that all recommendations produced by AI are verified against corporate principles, regulatory standards, and limits of safety before they are forwarded to the engineer to be reviewed finally (Al-Dushaishi and Al-Sofi, 2025).

The validation of the framework was done by two primary data collection methods, and the two methods produced strong and consistent results. The expert interview with a drilling specialist proved that the hierarchical offset selection logic is more sensible and explainable than the current manual processes, automatic risk warning, based on daily drilling reports, would come in handy, and the rules-based engine of governance would make the plans better, requiring less effort on the part of engineers to check compliance with rules on a routine basis and elevate the level of their judgment. The other highly important structural recommendation given by the expert was the fact that the appropriate flow must make the governance engine precede the engineer review and not a parallel process, as was incorporated in the final framework design. The survey of 22 drilling engineers who responded to the survey was also used to support these findings as they concurred on all the ten questions about 78 percent on average. The greatest support

with 81.8% agreement was automatic risk extraction, AI-generated parameter recommendations, safety rule verification and the concept of the framework. These findings affirm that the framework is able to fill veritable and well-known gaps in existing planning processes (Rahman, M., Ahmed, S., and Khan, R.). (2023)).

The paper also confirms that the future of the drilling program preparation is in the combination of artificial intelligence and human engineering judgment. The framework will not replace the drilling engineer but empower them with a traceable, data-driven and structured toolset. The framework offers the complete benefit of the available experience in history by systemising the experience of past operations to such that it may be readily available and memorable. This is in concurrence with the greater aim of digital transformation in the energy sector that is to have safer, efficient and predictable well delivery processes. As the industry advances and wells become increasingly complex and costly, organisations will need conceptual integrated models of the kind in order to continue delivering quality in their planning and reduce recidivism failures as well as remain competitive in a progressively data-driven operational environment.

REFERENCES

[1] Adjei, K. Y., Odor, G. E., Nurein, S. A., Ogu-Opara, P. C., Ugwu, O. C., Owunna, I. B., & Virginia, E. O. (2025). Optimizing well placement and reducing costs using AI-driven automation in drilling operations. *World Journal of Advanced Research and Reviews*, *25*(2), 1029–1038.

[2] ADNOC. (2024). *ADNOC and AIQ accelerate deployment of industry-first AR360 AI solution*. https://www.adnoc.ae/en/news-and-media/press-releases/2024/adnoc-and-aiq-accelerate-deployment-of-industry-first-ar360-ai-solution

[3] Al-Dushaishi, M. F., & Al-Sofi, H. (2025). Drilling optimization using artificial neural networks and field data from southern Iraq. *Energies*, *9*(2), 37. https://doi.org/10.3390/en09020037

[4] Allawi, R. H., Al-Mudhafar, W. J., Abbas, M. A., & Wood, D. A. (2025). Leveraging boosting machine learning for drilling rate of penetration (ROP) prediction based on drilling and petrophysical parameters. *Artificial Intelligence in Geosciences*, 100121.

[5] Al-Mudhafar, W. J., Wood, D., Al-Obaidi, D., & Wojtanowicz, A. (2023). Well placement optimization through the triple-completion gas and downhole water sink-assisted gravity drainage (TC-GDWS-AGD) EOR process. *Energies*, *16*(4), 1790.

[6] Al-Qattan, A., & Al-Shammari, M. (2023). *Learning from offset wells to optimize drilling programs in carbonate reservoirs* [Paper presentation]. SPE Middle East Oil & Gas Conference. https://doi.org/10.2118/214101-MS

[7] Al-Rubaii, M., Al-Shargabi, M., & Al-Shehri, D. (2023). Hole cleaning during drilling oil and gas wells: A review for hole-cleaning chemistry and engineering parameters. *Advances in Materials Science and Engineering*, *2023*(1), Article 6688500.

[8] Amani, M., et al. (2024). Challenges in automated drilling program generation. *Petroleum Exploration and Development*, *51*(2), 340–355.

[9] Amani, M., et al. (2024). *Intelligent well planning: AI-supported decision workflows for complex wells* [Paper presentation]. SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/215004-MS

[10] Aneke, A., & Eteyen, O. E. (2024). Enhancing operational efficiency in Nigerian oil exploration: The impact of real-time monitoring technologies on non-productive time. *Journal of Sustainable Development of Transport and Logistics*, *9*(2), 43–52.

[11] Arinze, C. A., Izionworu, V. O., Isong, D., Daudu, C. D., & Adefemi, A. (2024). Integrating artificial intelligence into engineering processes for improved efficiency and safety in oil and gas operations. *Open Access Research Journal of Engineering and Technology*, *6*(1), 39–51.

[12] Baquero Rico, N. C. (2021). Risk assessment for blowouts and kicks in oil and gas wells during drilling activities: A review [Unpublished manuscript or Thesis].

[13] Bazeley, P., & Jackson, K. (2013). *Qualitative data analysis with NVivo*. Sage Publications.

[14] Black, J., & Murray, A. D. (2019). Regulating AI and machine learning: Setting the regulatory agenda. *European Journal of Law and Technology*, *10*(3).

[15] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

[16] Bryman, A. (2016). *Social research methods*. Oxford University Press.

[17] Chen, F., Sun, L., Jiang, B., Huo, X., Pan, X., Feng, C., & Zhang, Z. (2025). A review of AI applications in unconventional oil and gas exploration and development. *Energies*, *18*(2), 391.

[18] Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches. Sage Publications.

[19] Deziel, N. C., Clark, C. J., Casey, J. A., Bell, M. L., Plata, D. L., & Saiers, J. E. (2022). Assessing exposure to unconventional oil and gas development: Strengths, challenges, and implications for epidemiologic research. *Current Environmental Health Reports*, *9*(3), 436–450.

[20] Díaz, E., & Spagnoli, G. (2024). Gradient boosting trees with Bayesian optimization to predict activity from other geotechnical parameters. *Marine Georesources & Geotechnology*, *42*(8), 1075–1085.

[21] Elekhteiar, M., et al. (2025). Reduction of non-productive time through digital well planning. *SPE Drilling and Completion*, *40*(1), 88–102.

[22] Epelle, E. I., & Gerogiorgis, D. I. (2020). A review of technological advances and open challenges for oil and gas drilling systems engineering. *AIChE Journal*, *66*(4), e16842.

[23] Feng, G. Y., & Han, J. X. (2015). The oilfield production prediction model based on principal component analysis and least squares support vector machine. *Computer Knowledge and Technology*, *11*(31), 144–147.

[24] Feng, Y., & Han, Z. (2015). Hierarchical well similarity analysis for offset selection. *Journal of Natural Gas Science and Engineering*, *27*, 1535–1545.

[25] Gao, W., & Sun, R. (2023). Knowledge-driven drilling recommendations using case-based reasoning and operational history. *Engineering Applications of Artificial Intelligence*, *122*, 106078. https://doi.org/10.1016/j.engappai.2023.106078

[26] Hanif, H. R. (2024). The role of artificial intelligence in optimizing oil exploration and production. *Eurasian Journal of Chemical, Medicinal and Petroleum Research*, *3*(5), 176–190.

[27] Haouel, C., & Nemeslaki, A. (2024). Digital transformation in oil and gas industry: Opportunities and challenges. *Periodica Polytechnica Social and Management Sciences*, *32*(1), 1–16.

[28] Hassanvand, M., Moradi, S., Fattahi, M., Zargar, G., & Kamari, M. (2018). Estimation of rock uniaxial compressive strength for an Iranian carbonate oil reservoir: Modeling vs. artificial neural network application. *Petroleum Research*, *3*(4), 336–345.

[29] Hersum, T. (2023). *Subsurface data and AI in Equinor*. Equinor. https://www.equinor.com/energy/data-in-equinor

[30] Ibrahim, M. O. (2004). *Computer application-electro mechanical systems* (Doctoral dissertation). Dublin City University.

[31] Johnston, M. P. (2017). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, *3*(3), 619–626.

[32] Khan, Z. (2025). Automated drilling process control using artificial intelligence: A novel framework for unconventional reservoirs and fracking operations. *International Journal of Engineering Research & Technology (IJERT)*, *14*(12), 1–15.

[33] Koroteev, D., & Tekic, Z. (2021). Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios. *IEEE Access*, *9*, 163107–163119.

[34] Mertiri, S. (2025). Artificial intelligence in project selection in process industries: Enhancing prioritization, risk management, and decision-making [Master's thesis, Stockholm University]. DiVA Portal.

[35] Mohammadinia, F., et al. (2025). Rate of penetration prediction in drilling operations via machine-learning frameworks. *Journal of Petroleum Science and Engineering*, *6*(2). https://doi.org/10.1007/s13202-025-02020-9

[36] Ochoa, D. (2024). Drilling automation and innovation. *Journal of Petroleum Technology*, *76*(2). https://jpt.spe.org/drilling-automation-and-innovation-2024

[37] Page, M. J., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71.

[38] Petrofac. (2024). *Digital well delivery and data-driven planning*. Petrofac Technical Paper Series. https://www.petrofac.com
[39] Póvoas, M. S., Moreira, J. F., Martins Neto, S. V., Carvalho, C. A. S., Cezario, B. S., Guedes, A. L. A., & Lima, G. B. A. (2025). Artificial intelligence in the oil and gas industry: Applications, challenges, and future directions. *Applied Sciences*, *15*(14), 7918.
[40] Prestidge, K. L. (2022). *Digital transformation in the oil and gas industry: Challenges and potential solutions* (Doctoral dissertation). Massachusetts Institute of Technology.
[41] Rahman, M., Ahmed, S., & Khan, R. (2023). *Challenges in conventional drilling program preparation workflows* [Paper presentation]. SPE Asia Pacific Oil and Gas Conference, Paper SPE-215234.
[42] Rahman, M., Ahmed, S., & Khan, R. (2023). Machine-learning-assisted drilling performance prediction and decision support. *SPE Journal*, *28*(2), 455–468. https://doi.org/10.2118/212345-PA
[43] Salem, A. M., Yakoot, M. S., & Mahmoud, O. (2022). Addressing diverse petroleum industry problems using machine learning techniques: Literary methodology—Spotlight on predicting well integrity failures. *ACS Omega*, *7*(3), 2504–2519.
[44] Salem, A., et al. (2022). Systematic failure mapping from offset well operations. *Reliability Engineering and System Safety*, *223*, 108–120.
[45] Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students*. Pearson Education.
[46] Shakirov, A., & Xie, Y. (2024). Machine learning-based drilling parameter prediction for real-time well control. *SPE Drilling & Completion*. https://doi.org/10.2118/214562-PA
[47] Shokouhi, S. V., Skalle, P., & Aamodt, A. (2014). An overview of case-based reasoning applications in drilling engineering. *Artificial Intelligence Review*, *41*(3), 317–329.
[48] Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in oil and gas industry. *Petroleum Research*, *6*(4), 379–391.
[49] SLB. (2025). *Eni using digital, AI, and automation to drill beyond limits*. https://www.slb.com/resource-library/case-study-with-navigation/di/eni-using-digital-ai-and-automation-to-drill-beyond-limits
[50] Teixeira, A., & Secchi, A. (2019). Machine learning models to support reservoir production optimization. *IFAC-PapersOnLine*, *52*(1), 498–501.
[51] Vishnumolakala, N. (2022). Downhole intelligence for drilling systems using supervised and deep reinforcement learning techniques (Doctoral dissertation).
[52] Wang, Z., Li, S., Jin, Z., Li, Z., Liu, Q., & Zhang, K. (2023). Oil and gas pathway to net-zero: Review and outlook. *Energy Strategy Reviews*, *45*, 101048.
[53] Williams, G., Meisel, N. A., Simpson, T. W., & McComb, C. (2022). Design for artificial intelligence: Proposing a conceptual framework grounded in data wrangling. *Journal of Computing and Information Science in Engineering*, *22*(6), 060903.

Algorithmic Trust and Choice Liquidity: A Dual-Pathway Framework for AI-Mediated Consumer Decision Architecture

Dr. Markus Rach
Canadian University Dubai
markus.rach@cud.ac.ae

Abstract—This conceptual paper proposes a dual-pathway framework explaining how artificial intelligence recommender systems reshape consumer brand evaluation and choice behavior. Integrating trust transfer theory, brand equity frameworks, and consideration-set research, two interacting pathways are proposed, a traditional brand-equity route and a novel AI-mediated route, and choice liquidity is introduced as a structural construct capturing the algorithmically driven reconfiguration of consideration sets across purchase episodes. Choice liquidity is formally differentiated from consideration-set volatility, brand switching, and exploration in recommender systems. Seven hypotheses address trust transfer, substitution and amplification effects, and contextual moderation. A compressed longitudinal design is outlined as the vehicle for future empirical evaluation.

Keywords—AI recommender systems, trust transfer, brand equity, consideration sets, choice liquidity, algorithmic mediation

I. INTRODUCTION

A. Context and Problem

Artificial intelligence (AI) recommender systems (RS) have become structural features of digital commerce, operating across e-commerce, media streaming, banking, and social platforms. By learning from behavioral data, they generate personalized recommendations that reduce consumer search costs. Empirical research demonstrates that when consumers perceive an RS as accurate, transparent, and personalized, trust in the system transfers to recommended products and brands, a mechanism grounded in trust transfer theory [1].

Simultaneously, branding theory conceptualizes brand equity as encompassing awareness, associations, and loyalty [2], a cognitive shortcut that facilitates decision-making and mitigates perceived risk [3]. Traditional consumer decision models assume a direct route from brand equity to purchase outcomes, a pathway designated as Path 1 in this paper's framework. The emergence of AI-mediated recommendation introduces a parallel route: trust in the algorithmic system itself may transfer to, substitute for, or amplify pre-existing brand equity.

B. Research Gaps

Existing studies examine immediate post-recommendation trust outcomes without integrating AI trust and brand equity into a unified theoretical model [4], [6]. The longitudinal consequences of repeated algorithmic mediation on brand loyalty, switching, and market diversity remain theoretically underspecified. Evidence of echo-chamber effects in AI RS has not been integrated into branding or equity frameworks.

No construct captures the structural reconfiguration of consideration-set membership across purchase episodes, leaving a gap between single-episode trust studies and longitudinal market-diversity outcomes.

C. Contributions

This paper proposes three contributions to the existing body of knowledge. First, it conceptualizes a dual-pathway framework integrating trust transfer theory, brand equity, and consideration-set research. Second, it introduces choice liquidity as a structural property of AI-mediated decision environments. Third, it advances a compressed longitudinal experimental design as the empirical vehicle for addressing the absence of validated methods for capturing dynamic AI RS effects across consideration sets and brand equity simultaneously.

II. LITERATURE REVIEW

A search via consensus. app across 170 million papers yielded 1,130 results; 607 were screened, 351 met eligibility criteria, and 50 were included across eight targeted search groups covering trust transfer, consideration-set effects, brand equity, and echo-chamber dynamics.

A. Trust Transfer from AI to Brands

Trust transfer theory suggests that trust generalizes from one entity to a related entity when a meaningful connection is perceived [1]. Studies confirm positive spillover from AI recommender trust to brand trust; among Generation Z consumers, for instance, AI accuracy evaluations significantly enhanced brand trust and purchase decisions [4]. Voice assistant research confirms that system trust mediates shopping satisfaction through consideration-set adjustment [5].

Field experiments using causal mediation confirm that RS increase both the size and depth of actively considered options [6].

B. Brand Equity and Consideration Sets

Brand equity accrues through awareness, associations, and loyalty, reducing perceived risk and increasing the probability that a brand enters a consumer's consideration set [2]. The literature understands consideration sets as memory-based structures shaped by prior experience and marketing communications [7]. AI RS disrupt this architecture by

foregrounding options which are outside consumers' awareness [6].

While AI can expand consideration sets by surfacing novel alternatives, unconditional trust in AI leads consumers to default to recommended options, narrowing choice diversity. This echo-chamber dynamic, where high algorithmic trust reinforces rather than broadens consumption patterns [8], remains unaddressed in branding and equity frameworks.

C. Contextual Moderators

For hedonic products, human recommenders outperform AI due to mentalizing activation; for utilitarian purchases or with anthropomorphic cues, AI is comparably persuasive [9]. Consistent with dual-process theory, high-stakes decisions increase reliance on accuracy and transparency signals. Brand awareness has been shown to moderate RS trust transfer, with lower awareness amplifying and higher awareness dampening the effect [10].

Transparency and explainability emerge as critical levers: explainable AI demonstrably increases user confidence and calibrates reliance [11]. Trust calibration, the degree to which consumers regulate algorithmic guidance, moderates transfer magnitude and consideration-set composition [11].

D. Research Gaps and Theoretical Need

Three interconnected gaps emerge. First, no unified framework integrates trust transfer and consideration-set effects longitudinally. Second, while brand awareness has been tested as a moderator of AI RS effects [10], no framework models trust-level substitution versus amplification across the full range of brand equity levels. Third, no construct captures the structural reconfiguration of consideration sets under sustained algorithmic mediation. A fourth, methodological gap underlies all three: no validated design exists for isolating AI RS trust effects longitudinally across consideration sets and brand equity within a single study. These gaps collectively motivate the dual-pathway framework, the choice liquidity construct, and the compressed longitudinal design introduced in this paper.

III. CONCEPTUAL FRAMEWORK AND HYPOTHESES

A. Choice Liquidity: Definition and Differentiation

Choice liquidity is defined as the algorithmically driven reconfiguration of consideration-set membership across sequential purchase episodes, a latent structural property of the consumer's decision environment under sustained AI mediation.

Three distinctions to existing constructs are critical. First, although consideration-set volatility can encompass both size and compositional change, it remains a descriptive, episode-level observation agnostic to cause. Choice liquidity attributes the reconfiguration specifically to algorithmic mediation: membership shifts are structurally governed by repeated AI recommendation across a decision sequence, not by random search or volitional change. Consideration-set volatility is therefore a necessary but insufficient condition for choice liquidity.

Second, choice liquidity differs from brand switching: switching is a discrete transaction-level behavior; choice liquidity is the latent environmental condition that may or may not produce switching. A consumer can exhibit high choice liquidity, cycling through varied brand exposures across episodes, while maintaining identical final purchase behavior.

Third, choice liquidity differs from exploration in RS: exploration is a system-side design objective (algorithmic diversity injection); choice liquidity is the consumer-side outcome. The same exploratory algorithm produces high or low choice liquidity depending on whether the consumer's trust is calibrated or unconditional. This carries direct policy implications: welfare assessments must account for consumer-side response, not algorithmic intent alone.

Choice liquidity is bidirectional: expansion (new brands surfaced under high-transparency, accurate AI signals) and narrowing (reduced set diversity through default reliance under echo-chamber conditions). This bidirectionality distinguishes it from scalar quantity measures such as consideration-set size.

B. Dual-Pathway Framework

This paper proposes a dual-pathway framework (Fig. 1) with two interacting routes:

Path 1 – Brand-Equity Route (Traditional): Brand equity → purchase intention [2].

Path 2 – AI-Mediated Route: AI trust → brand trust (via trust transfer) → dynamic consideration set → choice liquidity → brand switching/loyalty.

Brand trust serves as the proximal driver of purchase outcomes in both routes: in Path 1 it operates through established brand associations [2], [3]; in Path 2 it is generated explicitly through algorithmic trust transfer [1]. The pathways interact through brand equity: high-equity brands experience amplification; low-to-medium equity brands experience substitution. Product type, perceived risk, anthropomorphism, transparency, and trust calibration moderate the relative strength of each pathway.

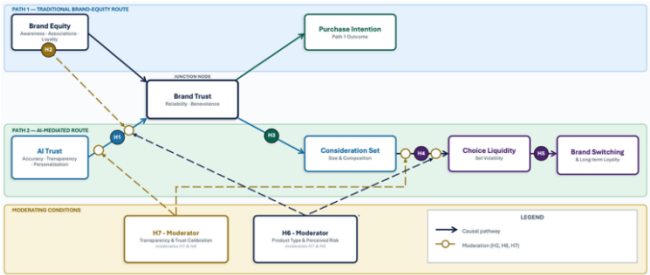


Fig. 1. Dual-Pathway Framework for AI-Mediated Consumer Decision Architecture. Solid arrows = causal paths; dashed arrows = moderation; ○ = moderation intersection points. In Path 1, brand trust operates implicitly through established brand associations; in Path 2 it is generated explicitly through trust transfer.

C. Hypotheses

Several hypotheses have prior empirical support. However, given the low replication rate in marketing studies [12], where prior empirical work exists, the proposed hypotheses also serve a replication purpose.

H1 (Trust Transfer): Trust in AI RS positively influences trust in the recommended brand.

H2 (Substitution vs. Amplification): The positive effect of AI trust on brand trust is stronger for low- and medium-equity brands (substitution) than for high-equity brands (amplification).

H3 (Consideration Set Expansion): Higher AI trust increases consideration-set size, mediated by trust in the recommended brand.

H4 (Choice Liquidity): Repeated reliance on AI RS increases choice liquidity, defined as volatility in consideration-set membership, which manifests as both expansion and echo-chamber narrowing.

H5 (Switching and Loyalty): Elevated choice liquidity increases short-term brand switching without reducing long-term category loyalty. H5 is proposed for future longitudinal research.

H6 (Moderation by Product Type and Risk): The substitution effect (H2) and choice liquidity (H4) are stronger for utilitarian products and high-risk purchase contexts than for hedonic products and low-risk contexts.

H7 (Trust Calibration): Transparency, explainability, anthropomorphism and user control moderate trust transfer and choice liquidity; calibrated trust reduces echo-chamber narrowing and promotes informed exploration.

IV. RESEARCH DESIGN

A. Design Overview and Novel Contribution

A compressed longitudinal experiment captures dynamic changes in trust, consideration sets, and brand switching across repeated decision episodes without multi-year deployment. Three innovations distinguish this design: (1) purchase occasions are simulated within one week, enabling longitudinal causal inference; (2) AI trust signals are manipulated experimentally, permitting clean identification of trust-pathway effects; and (3) validated survey scales, behavioral set-membership tracking, and information-theoretic diversity metrics are integrated within a single design.

Three hundred participants, stratified by age, gender, and digital literacy, are randomly assigned to conditions, satisfying thresholds for stable SEM estimation [13] and multilevel model power [14]. Although mediation and moderation are tested via regression-based path analysis [19], the structural complexity of the multi-level, multi-wave design, encompassing latent growth models and cross-level interactions, justifies the 300-participant threshold recommended for complex structural models [13]. Compressed simulation designs have demonstrated validity for capturing dynamic consumer choice [15]; participants therefore complete four purchase episodes at 48-hour intervals over one week.

B. Stimuli and Manipulations

Stimuli consist of two product categories, utilitarian (household consumables) and hedonic (premium confectionery), presented through a custom mock e-commerce interface (Fig. 2). The embedded AI engine varies three trust signal dimensions between subjects: transparency (explainable versus black-box), anthropomorphism (human-like avatar versus neutral interface), and stated accuracy (high precision versus unspecified). The three dimensions are administered as a composite condition, all signals present (high trust) versus all signals absent (low trust), yielding two trust-signal groups crossed with the calibration prompt factor to produce four experimental conditions of approximately 75 participants each.

Brand equity is manipulated through real and fictitious brand pairings, pretested for awareness and favorability. Half the sample receives trust calibration prompts; the other half serves as controls.

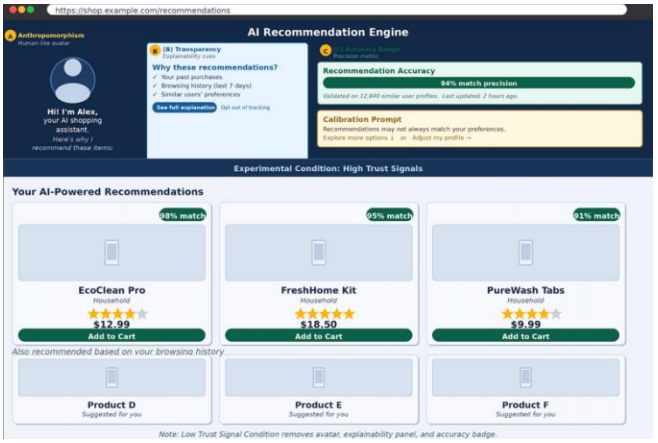


Fig. 2. Prototype Mock E-Commerce Interface (High Trust Signal Condition). Callouts (A)–(C): (A) anthropomorphism, (B) transparency/explainability, (C) accuracy badge. The low trust condition removes all three.

C. Construct Measurement

AI trust is measured using a validated multi-item scale [16]. Brand trust is assessed using an established consumer-brand trust scale [17]. Consideration-set size and membership are captured behaviorally as self-reported candidate sets per episode.

Choice liquidity is operationalized as: (1) within-subject variance in consideration-set membership across episodes; (2) proportion of novel brands introduced per episode; and (3) Shannon entropy of brand distributions, collectively distinguishing exploration from echo-chamber consolidation. The indirect effect of AI trust on consideration-set size via brand trust (H3) is tested using bootstrapped confidence intervals within the regression-based path analysis framework

[19]. Multilevel mixed-effects models accommodate the repeated-measures structure [18]; regression-based path analysis tests the framework's hypothesized mediation and moderation paths [19]; latent growth models capture trajectories across episodes [20].

V. CONCLUSION

AI RS reshape consumer decision architecture by introducing an algorithmically governed trust pathway alongside traditional brand equity mechanisms. This paper proposes a dual-pathway framework integrating trust transfer theory, brand equity, and consideration-set research, and introduces choice liquidity, a formally differentiated construct capturing consideration-set reconfiguration under algorithmic mediation.

Seven testable hypotheses and a compressed longitudinal design provide a rigorous vehicle for future empirical evaluation. The framework bridges marketing, behavioral economics, and platform governance by treating AI as a structural feature of choice environments. For practitioners, transparency and calibration mechanisms are prerequisites for sustainable AI-mediated brand strategy. For policymakers, the model supports regulatory interest in algorithmic transparency, user control, and disclosure standards. Future research should implement the design across diverse cultural contexts and examine the interplay between AI recommendation and complementary trust mechanisms such as influencer endorsement.

REFERENCES

[1] D. Belanche, L. V. Casaló, C. Flavián, and J. Schepers, "Trust transfer in the continued usage of public e-services," Inf. Manag., vol. 51, no. 6, pp. 627–640, 2014.

[2] D. A. Aaker, "The value of brand equity," J. Bus. Strategy, vol. 13, no. 4, pp. 27–32, 1992.

[3] T. Erdem, J. Swait, S. Broniarczyk, D. Chakravarti, J.-N. Kapferer, M. Keane, J. Roberts, J.-B. E. M. Steenkamp, and F. Zettelmeyer, "Brand equity, consumer learning and choice," Mark. Lett., vol. 10, no. 3, pp. 301–318, 1999.

[4] C. R. Guerra-Tamez, K. Kraul Flores, G. M. Serna-Mendiburu, D. Chavelas Robles, and J. Ibarra Cortés, "Decoding Gen Z: AI's influence on brand trust and purchasing behavior," Front. Artif. Intell., vol. 7, 2024. https://doi.org/10.3389/frai.2024.1323512

[5] A. Mari and R. Algesheimer, "The role of trusting beliefs in voice assistants during voice shopping," in Proc. HICSS 2021. https://aisel.aisnet.org/hicss-54/in/ai_based_assistants/7

[6] X. Li, J. Grahl, and O. Hinz, "How do recommender systems lead to consumer purchases? A causal mediation analysis of a field experiment," Inf. Syst. Res., vol. 33, no. 2, pp. 620–637, 2022.

[7] P. Nedungadi, "Recall and consumer consideration sets: Influencing choice without altering brand evaluations," J. Consum. Res., vol. 17, no. 3, pp. 263–276, 1990.

[8] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, and Y. Zhang, "Understanding echo chambers in e-commerce recommender systems," in Proc. 43rd Int. ACM SIGIR Conf., 2020, pp. 2261–2270.

[9] A. H. Wien and A. M. Peluso, "Influence of human versus AI recommenders: The roles of product type and cognitive processes," J. Bus. Res., vol. 137, pp. 13–27, 2021.

[10] W. Chang and J. Park, "A comparative study on the effect of ChatGPT recommendation and AI recommender systems on the formation of a consideration set," J. Retail. Consum. Serv., vol. 78, p. 103743, 2024.

[11] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," Int. J. Hum.-Comput. Stud., vol. 146, p. 102551, 2021.

[12] H. Evanschitzky, C. Baumgarth, R. Hubbard, and J. S. Armstrong, "Replication research's disturbing trend," J. Bus. Res., vol. 60, no. 4, pp. 411–415, 2007.

[13] E. J. Wolf, K. M. Harrington, S. L. Clark, and M. W. Miller, "Sample size requirements for structural equation models," Educ. Psychol. Meas., vol. 73, no. 6, pp. 913–934, 2013.

[14] C. J. Maas and J. J. Hox, "Sufficient sample sizes for multilevel modeling," Methodology, vol. 1, no. 3, pp. 86–92, 2005.

[15] R. R. Burke, B. A. Harlam, B. E. Kahn, and L. M. Lodish, "Comparing dynamic consumer choice in real and computer-simulated environments," J. Consum. Res., vol. 19, no. 1, pp. 71–82, 1992.

[16] M. J. McGrath, O. Lack, J. Tisch, and A. Duenser, "Measuring trust in artificial intelligence: Validation of an established scale and its short form," Front. Artif. Intell., vol. 8, p. 1582880, 2025.

[17] E. Delgado-Ballester, J. L. Munuera-Aleman, and M. J. Yague-Guillen, "Development and validation of a brand trust scale," Int. J. Mark. Res., vol. 45, no. 1, pp. 35–53, 2003.

[18] S. W. Raudenbush and A. S. Bryk, Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed. Thousand Oaks: Sage, 2002.

[19] A. F. Hayes, Introduction to Mediation, Moderation, and Conditional Process Analysis, 3rd ed. New York: Guilford Press, 2022.

[20] P. J. Curran, K. Obeidat, and D. Losardo, "Twelve frequently asked questions about growth curve modeling," J. Cogn. Dev., vol. 11, no. 2, pp. 121–136, 2010.

Designing Deterministic AI Agents in Enterprise Platforms: A Schema-Guided Reasoning Approach

Ignatii Anokhin
Department of Business Informatics
National Research University Higher School of Economics
Moscow, Russia
igananokhin@edu.hse.ru

Abstract—The adoption of AI agents in enterprise systems is constrained by limited controllability, reproducibility, and transparency of their reasoning processes, which restricts their use in business-critical scenarios. This study aims to explore how reasoning in enterprise AI agents can be structured to support reliable multi-stage task execution. To address this, the paper investigates the application of Schema-Guided Reasoning (SGR), a structured approach that decomposes reasoning into stages with predefined schemas, validation checkpoints, and iterative control mechanisms. The contribution of the study is the validation of SGR applicability in an enterprise support scenario and the formulation of a reasoning workflow suitable for corporate environments.

Keywords—AI agents, deterministic reasoning, enterprise platforms, schema-guided reasoning, LLM orchestration, enterprise automation

I. INTRODUCTION

A. Context

Recent advances in large language models (LLMs) have enabled the development of autonomous AI agents capable of planning, contextual interpretation, and interaction with external systems. These agents are increasingly applied across domains, including enterprise environments, where the focus is shifting toward goal-oriented and proactive systems [1].

In enterprise platforms such as CRM, ERP, and xRM systems, AI agents offer significant potential for automating routine operations, supporting decision-making, analyzing data, and managing organizational resources [2]. However, enterprise deployment introduces stricter requirements compared to consumer use cases. Key expectations include predictable behavior, reproducibility, controllability of reasoning, compliance with internal policies, auditability, and secure handling of sensitive data [2], [4].

Additionally, enterprise agents must operate within complex IT ecosystems, integrating with internal systems, tools, and data sources while maintaining governance and reliability [6], [12]. As a result, the challenge extends beyond output quality to the organization and control of the reasoning process itself.

B. Problem Statement

Despite growing adoption, enterprise implementation of AI agents faces several persistent challenges.

First, reasoning processes are often non-deterministic: similar inputs may produce different outputs, limiting their applicability in business-critical workflows [2], [17].

Second, multi-step scenarios require handling large and evolving contexts, while LLMs are constrained by context window limitations and may fail to retain relevant information across reasoning steps [1], [13].

Third, many agent systems are tightly coupled with specific models, prompting strategies, and tool-calling mechanisms, reducing architectural portability and complicating long-term evolution [1], [12], [13].

Fourth, enterprise constraints frequently necessitate the use of local or smaller models due to security, cost, and infrastructure requirements. However, such models typically demonstrate weaker reasoning capabilities, especially in complex multi-step tasks [3], [9], [15].

Finally, existing approaches often lack sufficient traceability of intermediate reasoning steps, limiting testing, debugging, verification, and audit capabilities [2], [5], [17].

These limitations are also reflected in empirical benchmarks. For example, CRMArena demonstrates that even advanced LLM agents struggle with realistic CRM tasks in professional environments [8].

C. Research Gap

Prior research proposes various techniques to improve agent reliability, including structured planning, context management, intermediate validation, governed execution, and separation of architectural logic from the underlying model [5], [7], [13], [17]. However, these approaches remain fragmented and do not form a unified, widely adopted framework for structuring reasoning in enterprise agents.

In practice, organizations must combine multiple techniques and validate them independently within specific business scenarios [1], [2]. This creates a gap between available architectural concepts and the need for a coherent mechanism that enables reasoning to function as a controlled, reproducible, and verifiable process in complex enterprise workflows.

D. Research Question

This study addresses the following research question: How can reasoning processes in enterprise AI agents be structured to enable reliable multi-stage task execution?

A related sub-question examines how such structuring can support determinism, reproducibility, model-agnostic operation, and traceability of reasoning.

E. Contributions

This paper investigates the application of a structured reasoning control mechanism in an enterprise scenario. The approach aims to make multi-step reasoning explicit, decomposed, and verifiable.

The study contributes by:

- demonstrating how structured reasoning can be applied in a real enterprise use case,
- defining a reasoning workflow aligned with enterprise requirements,
- and evaluating its applicability and potential effectiveness based on initial industrial validation.

Importantly, the contribution lies in the application and adaptation of existing but not widely accepted principles of structured reasoning, rather than proposing a new framework.

II. METHODOLOGY

A. Research Design

This study follows the Design Science Research (DSR) methodology, focusing on the development and practical validation of an architectural artifact addressing a real-world problem. The research is conducted within a Russian enterprise vendor of an xRM low-code platform used for ITSM solutions. This setting enables the investigation of AI agents in realistic operational contexts, including incident handling, user request processing, knowledge base utilization, etc. Consequently, the proposed artifact is designed with direct applicability to production environments.

The research process consists of several stages. First, key limitations of existing agent solutions are identified and formalized. Second, relevant architectural approaches are analyzed. Third, a target reasoning design is developed and applied to a selected enterprise scenario. Finally, the proposed approach is evaluated through a controlled experiment.

B. Problem Identification

Problem identification was based on analyzing existing AI agents deployed within the company's products and collecting feedback from key stakeholders. Three groups participated: product managers, engineers responsible for configuring agent scenarios, and end users.

The study included 10 interviews (4 internal teams and 6 customer teams) and analysis of acceptance testing results, including expert evaluations of scenario performance. This enabled the construction of a representative set of enterprise automation scenarios and identification of recurring limitations.

The most common scenarios were related to user support automation, including ticket classification, operator copilots for solution search, report generation, incident pattern detection, and monitoring data processing. Detailed feedback was collected for each scenario, including specific failure cases and instability patterns.

The analysis revealed that the identified issues are systemic and consistent across scenarios. These findings align with the limitations described in Section 3, including non-deterministic reasoning, model dependency, and limited traceability.

A support copilot scenario for solution retrieval was selected as the primary evaluation case. This choice is justified by its high business impact, availability of quantitative feedback (current quality score: 6.8/10), and its representation of a typical multi-step reasoning process involving retrieval, analysis, and synthesis.

C. Knowledge Exploration

To determine an appropriate architectural approach, existing research on AI agents and structured reasoning was reviewed.

LLM agent architectures. Prior work shows that LLM agents typically consist of planning, memory, and execution components and are applied across diverse domains, including enterprise systems [1]. However, enterprise adoption highlights challenges related to reliability and reproducibility [2], while many architectures emphasize communication and tool integration [6].

Structured reasoning and controlled generation. Schema-aware generation and grammar-constrained decoding improve output consistency and reduce errors by enforcing predefined structures [17]. Frameworks such as Routine introduce step-wise planning and validation mechanisms to enhance robustness [9].

Schema-Guided Reasoning (SGR). SGR organizes reasoning through predefined schemas, intermediate validation, and structured outputs [19]. Prior studies indicate improvements in reproducibility, debuggability, and testing, although the approach requires careful schema design [18], [19].

Tool orchestration and coordination. Enterprise-oriented architectures propose blueprint-based orchestration and separation of execution logic from language models [12], [17]. The POLARIS framework highlights the role of typed planning and controlled execution in automation scenarios [7].

Context management. Approaches such as SagaLLM and DataLab emphasize structured context handling, validation, and state management across reasoning stages [13], [14], underscoring the importance of managing intermediate reasoning states.

D. Solution Design

The solution design followed a structured process.

First, the existing assistant workflow (AS IS) was analyzed in detail, including processing steps, system components, and decision points. This allowed formalizing the implicit reasoning process.

Second, failure points identified during problem analysis were examined, with a focus on context loss, uncontrolled generation, and absence of intermediate validation.

Based on the literature review, a target approach was selected that decomposes reasoning into structured stages with explicit control over inputs and outputs. The design emphasizes the ability to introduce this mechanism without modifying the core platform, using an external orchestration layer.

As a result, a target reasoning workflow (TO BE) was defined, separating reasoning, retrieval, and response generation into distinct, controlled stages.

E. Evaluation

The proposed approach is evaluated using an A/B testing methodology. Two assistant implementations are compared: the baseline (AS IS) and the structured approach (TO BE). Both operate on the same dataset and use the same language model, isolating the impact of architectural changes.

The dataset consists of 100 test cases representing real user requests for solution retrieval from knowledge bases and ticket history. Outputs are evaluated by ITSM product experts using a 10-point scale.

Evaluation metrics include average quality score, reproducibility, performance on small language models (SLM), stability across different models, and token consumption. This set of metrics enables assessment not only of answer quality but also of robustness, portability, and cost efficiency.

III. PRELIMINARY RESULTS

A. Existing Workflow

The current assistant workflow (AS IS) follows a standard retrieval-augmented generation (RAG) architecture widely used in enterprise applications. A user query is first encoded into a vector representation, which is then used to retrieve relevant documents from a vector database. Retrieved results are re-ranked and combined with the original query during prompt construction, after which the language model generates the final response.

Fig. 1. AS IS Workflow

While this approach enables basic retrieval and answer generation, it does not provide explicit control over the reasoning process. Decision-making logic remains encapsulated within the language model, limiting transparency, controllability, and adaptability.

The component structure includes an embedding model (USER-bge-m3), a vector search engine (pgvector), a reranker (colbert-xm), and a generation model (saiga_llama3_8b). The main limitations of this pipeline are the lack of task structuring, sensitivity to query formulation, and non-deterministic outputs.

TABLE I. AS IS PIPELINE COMPONENTS

№	Component	Function	Tool / Model	Limitations
1	Embedder	Encode user query into vector	USER-bge-m3	No task structure awareness
2	Retriever	Retrieve relevant documents	pgvector	Sensitive to query wording
3	Reranker	Improve document relevance	colbert-xm	Partial noise filtering
4	Prompt Constructor	Combine query and context	Prompt template	
5	Answer Generator	Generate final response	saiga_llama3_8b	Non-deterministic output

B. Analysis of Limitations

The analysis of the existing workflow highlights several critical issues related to the absence of explicit reasoning control.

First, the input task is processed in an unstructured form, leading to ambiguous interpretation of user requests. This is particularly problematic in support scenarios, where a single case may involve multiple interrelated issues.

Second, retrieval relies on one or a limited number of queries derived directly from the input text. As a result, retrieval quality depends heavily on wording rather than the underlying problem structure.

Third, there is no mechanism for validating intermediate results. The system proceeds directly to response generation without verifying the completeness or correctness of retrieved information, increasing the likelihood of hallucinations.

Fourth, the reasoning process is not traceable. It is not possible to reconstruct which data sources were used or how the final response was derived, which complicates debugging, testing, and auditing.

Finally, the approach demonstrates limited robustness when applied to smaller models. As model quality decreases, both answer accuracy and reasoning capability degrade, since all logic is embedded within a single generation step.

C. Selection of Target Approach

To address these limitations, Schema-Guided Reasoning (SGR) is selected as the target approach for structuring reasoning [19].

The core principle of SGR is the use of structured outputs, where both intermediate and final results are represented through predefined schemas (e.g., JSON). This reduces variability, enables validation, and enforces consistency [19], [20].

At the process level, SGR relies on three key patterns:

- **Cascade**, which decomposes tasks into sequential subtasks with structured outputs;
- **Routing**, which dynamically selects execution paths based on intermediate results;
- **Cycle**, which enables iterative refinement (e.g., repeated retrieval or validation).

Their combination forms a controlled reasoning pipeline with explicit transitions between stages.

SGR aligns with enterprise requirements by supporting determinism, reproducibility, traceability, and model-agnostic design. It can be implemented as an external orchestration layer without modifying the underlying platform.

Compared to alternative approaches, SGR provides stronger formalization. Methods such as Chain-of-Thought and Tree-of-Thought rely on implicit textual reasoning without structured validation. Graph-of-Thought extends these ideas but introduces computational overhead without ensuring control. ReAct and Plan-and-Execute incorporate tool usage and planning but lack strict schema enforcement. Program-of-Thought improves precision through code-like reasoning but is limited to formalizable tasks. Multi-agent and hierarchical planning approaches increase architectural complexity without guaranteeing reproducibility.

In contrast, SGR enforces structured reasoning through schemas, typed data, and validation checkpoints, making it more suitable for enterprise scenarios requiring stability and auditability.

Existing studies indicate that schema-constrained generation reduces errors and improves consistency [19], while SGR-based implementations enhance reproducibility and testing capabilities [20], [21]. The contribution of this work lies in applying and adapting these principles to enterprise ITSM scenarios rather than proposing a new framework.

D. Proposed Workflow

The proposed reasoning workflow (TO BE) implements SGR principles as a multi-stage controlled pipeline with explicit decomposition and validation. Each stage operates on typed inputs and produces validated outputs, enabling deterministic transitions between stages.

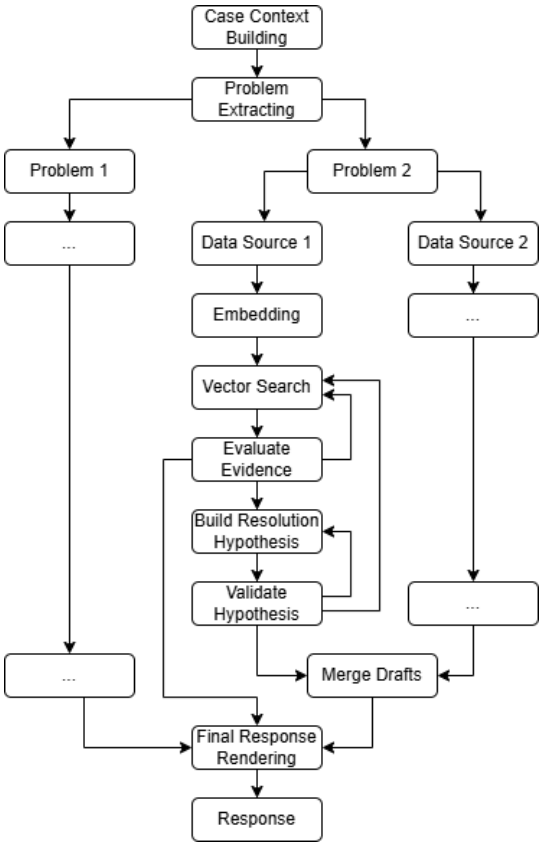


Fig. 2. TO BE Workflow

The process begins with Case Context Building, where the input is normalized into a structured representation. The Problem Extractor then applies cascade decomposition, splitting the case into independent structured problems (including symptoms, entities, and retrieval queries), reducing ambiguity.

Each problem is processed through a retrieval cycle. The Embedder converts structured queries into vector form, and the Retriever retrieves candidate evidence. The Source Evaluator applies routing and cycle mechanisms by assessing relevance and completeness, triggering additional retrieval if needed.

Based on validated evidence, the Resolution Hypothesis Builder generates structured solution hypotheses. The Resolution Hypothesis Evaluator verifies these hypotheses against the problem definition and schema, enabling feedback loops to earlier stages when inconsistencies are detected.

Validated hypotheses are combined in the Draft Builder, preserving the relationship between problems and solutions. Finally, the Final Response Renderer converts the structured draft into a user-facing response. This separation of reasoning and generation improves controllability and reproducibility.

Each component performs a distinct function within the reasoning process, eliminating the “black-box” effect and enabling traceability.

TABLE II. TO BE PIPELINE COMPONENTS

№	Component	Function	Tool / Model	Reasoning Role
1	Case Context Builder	Normalize input and assemble structured case context	saiga_llama3_8b	Reduces ambiguity before reasoning
2	Problem Extractor	Decompose case into one or more structured problems	saiga_llama3_8b	Makes reasoning explicit and decomposed
3	Embedder	Convert structured problem into vector representations for retrieval	USER-bge-m3	Enables semantic alignment between problem and knowledge
4	Retriever	Retrieve evidence using multiple queries and sources	pgvector	Provides initial evidence set
5	Source Evaluator	Filter and rank evidence, detect insufficiency (loop back to retrieval if needed, trigger response rendering if no relevant evidence found)	saiga_llama3_8b	Ensures evidence quality and relevance
6	Resolution Hypothesis Builder	Generate candidate solution based on structured problem and evidence	saiga_llama3_8b	Separates reasoning from final answer
7	Resolution Hypothesis Evaluator	Validate hypothesis against initial problem, evidence and schema (loop back to retrieval if needed)	saiga_llama3_8b	Enables intermediate validation and control
8	Draft Builder	Merge validated hypotheses into structured operator draft	saiga_llama3_8b	Preserves structure across subproblems
9	Final Response Renderer	Convert structured draft into final response	saiga_llama3_8b	Separates reasoning from final answer

Structured schemas play a central role in this design. Cases are represented as collections of problems with formalized attributes such as symptoms, entities, and retrieval queries.

Listing 1. Structured Problem Representation Output (JSON)

```
{
  "case_id": "INC-10452",
  "problems": [
    {
      "problem_id": "P1",
      "label": "VPN authentication failure",
      "symptoms": [
        "VPN login rejected",
        "issue started after password reset"
      ],
      "retrieval_queries": [
        "VPN login failed after password reset",
        "authentication error after credential change"
      ]
    }
  ]
}
```

Typed data models (e.g., using validation frameworks such as Pydantic) ensure consistency, reduce generation errors, and simplify integration between components [19], [20].

Listing 2. Typed Problem Representation (Python)

```
class Problem:
    problem_id: str
    label: str
    symptoms: list[str]
    retrieval_queries: list[str]

def extract_problems(case_text: str):
    """
    Extract structured problems
    using schema-guided LLM output
    """
    result = LLM(case_text,
                 schema=Problem)
    return result
```

E. MVP Limitations and Evaluation

An MVP implementation is developed using Python, with LangChain and LangGraph as the orchestration framework. Integration with the xRM platform is achieved via MCP, minimizing changes to the existing architecture.

A mid-sized local model (saiga_llama3_8b) is used as the baseline to evaluate whether structured reasoning can compensate for the limitations of smaller models.

Preliminary qualitative observations on a subset of cases indicate a reduction in irrelevant responses compared to the baseline RAG approach. A full quantitative evaluation is designed as an A/B test on 100 cases with expert scoring.

The current MVP is limited by partial platform integration, a restricted set of scenarios, and the need for manual schema design. Future work includes comprehensive quantitative evaluation using metrics such as answer quality, reproducibility, model robustness, and computational cost.

IV. CONCLUSION

This study examines the limitations of controllability and reproducibility in reasoning processes of enterprise AI agents, which restrict their use in business-critical contexts.

The paper explores the application of Schema-Guided Reasoning (SGR) as a mechanism for structuring reasoning through predefined stages, schemas, and validation checkpoints. The proposed workflow enables explicit task decomposition, intermediate validation, and separation of reasoning from response generation, supporting enterprise requirements such as traceability, determinism, and robustness.

Future work includes completing the MVP, conducting controlled A/B testing, and performing quantitative evaluation, followed by potential extension to production environments and other enterprise use cases.

REFERENCES

[1] T. Zhao, C. Ma, X. Feng, Z. Zhang, and H. Yang, “A survey on large language model based autonomous agents,” *Frontiers of Computer Science*, 2024, doi: 10.1007/s11704-024-40231-1.

[2] Muthusamy, Y. Rizk, K. Kate, P. Venkateswaran, and V. Ishakian, “Towards large language model-based personal agents in the enterprise: Current trends and open problems,” in *Proc. EMNLP Findings*, 2023, doi: 10.18653/v1/2023.findings-emnlp.461.

[3] Z. Wang et al., “A comprehensive survey of small language models in the era of large language models,” arXiv:2411.03350, 2024.

[4] Y. Yang, H. Chai, Y. Song, S. Qi, M. Wen, N. Li, et al., “A survey of AI agent protocols,” arXiv:2504.16736, 2025.

[5] Z. Moslemi, K. Koneru, Y.-T. Lee, S. Kumar, and R. Radhakrishnan, “POLARIS: Typed planning and governed execution for agentic AI in back-office automation,” arXiv, 2026.

[6] R. Shu, N. Das, M. Yuan, M. Sunkara, and Y. Zhang, “Towards effective GenAI multi-agent collaboration: Design and evaluation for enterprise applications,” arXiv, 2024.

[7] Y. Zeng et al., “Routine: A structural planning framework for LLM agent system in enterprise,” arXiv:2507.14447, 2025.

[8] K.-H. Huang, A. Prabhakar, S. Dhawan, Y. Mao, and H. Wang, “CRMArena: Understanding the capacity of LLM agents to perform professional CRM tasks in realistic environments,” arXiv:2411.02305, 2024.

[9] L. Tuggener et al., “So you want your private LLM at home? A survey and benchmark of methods for efficient GPTs,” in *Proc. Swiss Conf. Data Science*, 2024, doi: 10.1109/sds60720.2024.00036.

[10] Reitemeyer and H.-G. Fill, “Applying large language models in knowledge graph-based enterprise modeling: Challenges and opportunities,” arXiv:2501.03566, 2025.

[11] M. A. Ferrag, N. Tihanyi, and M. Debbah, “From LLM reasoning to autonomous AI agents: A comprehensive review,” arXiv:2504.19678, 2025.

[12] Kandogan, N. Bhutani, D. Zhang, L. Chen, and S. Gurajada, “Orchestrating agents and data for enterprise: A blueprint architecture for compound AI,” arXiv:2504.08148, 2025.

[13] Chang et al., “SagaLLM: Context management, validation, and transaction guarantees for multi-agent LLM planning,” *Proc. VLDB Endowment*, vol. 18, no. 5, pp. 1234–1247, 2025, doi: 10.14778/3750601.3750611.

[14] Weng et al., “DataLab: A unified platform for LLM-powered business intelligence,” arXiv:2412.02205, 2024.

[15] Vemulapalli, “Choosing the right model for enterprise AI applications: A comprehensive analysis of small language models versus large language models in enterprise architecture,” *European Modern Studies Journal*, vol. 9, no. 5, p. 25, 2025, doi: 10.59573/emsj.9(5).2025.25.

[16] Y. Liu et al., “Towards robust multi-modal reasoning via model selection,” arXiv:2310.08446, 2023.

[17] Qiu, Y. Ye, Z. Gao, X. Zou, and J. Chen, “Blueprint first, model second: A framework for deterministic LLM workflow,” arXiv:2508.02721, 2025.

[18] Cognia Lab, “Schema-Guided Reasoning: A new method of structuring reasoning in LLMs to reduce errors,” [Online]. Available: https://cognia-lab.com/schema-guidedreasoning

[19] R. Abdullin, “Schema-Guided Reasoning (SGR),” 2025. [Online]. Available: https://abdullin.com/schema-guided-reasoning/

[20] vamplabAI, “sgr-agent-core,” GitHub repository. [Online]. Available: https://github.com/vamplabAI/sgr-agent-core

[21] Redmadrobot, “Schema-Guided Reasoning: как научить языковые модели последовательно рассуждать,” Habr, [Online]. Available: https://habr.com/ru/companies/redmadrobot/articles/962236/

Chunk-Based Ensemble Simulation for AI Lifecycle Management

Mohammad Arif Ul Islam
Rochester Institute of Technology
Dubai, United Arab Emirates
ariful1251996@gmail.com

Dr. Ayman Ibrahim
Department of Graduate Programs & Research
Rochester Institute of Technology
Dubai, United Arab Emirates
ayman.ibrahim@rit.edu

Abstract—AI managers face a critical investment dilemma because validating dynamic model strategies requires longitudinal data, yet collecting such data demands the same expensive infrastructure that managers seek to validate. This paper presents a chunk-based ensemble framework that addresses this cold start problem by approximating the performance trajectory of periodic batch retraining using only static datasets. We partition existing data into chunks and train independent models on progressively combined segments, using ensemble averaging to approximate aggregate production behaviour under periodic retraining. We compare this approach against a full-data oracle baseline and a cumulative retraining baseline across seven algorithms. Validation demonstrated high prediction consistency under overlapping training data, alongside F1 improvements ranging from 0.16 to 0.52 percent for tree-based and distance-based algorithms, with results varying by algorithm family, though these gains should be interpreted with caution pending formal statistical validation. We discuss the distinction between our batch-update approximation and true incremental learning, situating the framework relative to concept drift and data stream literature. This domain-agnostic tool provides AI managers with preliminary evidence for infrastructure investment decisions without requiring longitudinal data collection.

Keywords—Dynamic Machine Learning, AI Lifecycle Management, Ensemble Methods, Batch Retraining Simulation, Cold Start Problem, Infrastructure Investment, Concept Drift.

I. INTRODUCTION

Organizations implementing production machine learning systems increasingly recognize that static models trained once and deployed indefinitely cannot maintain performance as operational data distributions evolve. Customer preferences shift, equipment degrades, and usage patterns emerge that were absent in historical training data. The theoretical solution of implementing dynamic models that continuously update as new data arrives faces a fundamental practical barrier.

AI managers encounter a chicken-and-egg dilemma. Validating whether adaptive systems justify their computational costs requires longitudinal data demonstrating how model performance evolves with sequential updates. Yet collecting such data demands the same expensive infrastructure, extended timeframes, and organizational resources that managers seek to validate before approving budgets. This creates the cold start problem in AI lifecycle management. The inability to make evidence-based decisions about dynamic modelling investments when working within conventional static dataset constraints [1].

The financial implications are substantial. Implementing dynamic systems requires investments in data pipelines, monitoring infrastructure, and computational resources. For mid-sized organizations, these costs can exceed hundreds of thousands of dollars annually, representing unacceptable risk without preliminary validation evidence.

A. Positioning Relative to Existing Dynamic Learning Approaches

Current machine learning literature offers sophisticated solutions for truly dynamic environments. Online learning algorithms such as Hoeffding Adaptive Trees [2] and ADWIN-based drift detectors [3] update incrementally with each new observation. Concept drift adaptation methods [4], [5] detect and respond to distribution shifts in data streams. Prequential or test-then-train evaluation protocols [6] provide rigorous assessment of stream learners by testing each model on new data before incorporating it into training. However, these approaches assume continuous data streams, real-time infrastructure, and the kind of longitudinal data that many organizations lack. They answer the question of how a deployed model should adapt to arriving data. Our framework addresses a different and earlier question like before committing to dynamic infrastructure, can an organization use its existing static dataset to estimate whether periodic retraining would yield measurable performance improvements? This distinction is critical. We do not claim to implement sequential or incremental learning. Rather, we propose a batch ensemble approximation, a simulation tool for investment decision support that operates under static data constraints.

B. Research Contribution

We present a chunk-based ensemble framework enabling AI managers to approximate the performance trajectory of periodic model retraining using only static datasets. The framework addresses three investment questions. First, would periodic updates provide measurable performance improvements? Second, would dynamic updating introduce prediction instability? Third, which algorithms benefit most from periodic retraining?

Unlike existing work focusing on developing sophisticated dynamic algorithms [7], [8] or online learners requiring stream infrastructure, our contribution is a practical

pre-investment validation tool for organizations working within conventional data constraints.

II. METHODOLOGY

A. Framework Design Philosophy and Scope

The chunk-based ensemble simulation framework operates on a fundamentally different principle than traditional temporal modelling. Rather than predicting how individual entities evolve over time, our framework evaluates the system-level question: what performance benefits emerge from periodically retraining production models as new data cohorts become available?

This aligns with real-world AI lifecycle scenarios where organizations receive discrete batches of records during enrollment periods, quarterly system extracts, or annual data updates rather than continuous real-time streams [9]. The framework is explicitly designed for the batch retraining setting in which an organization has an existing static dataset and wishes to estimate whether a policy of periodic retraining, as new cohorts arrive, would outperform a single static model trained once.

The framework does not model temporal dependencies between individual records, concept drift, or distribution shift. It does not implement incremental weight transfer between sequential models. Each chunk-trained model is independently initialized, making this an ensemble approximation rather than a true sequential learner. This distinction is acknowledged as both a design choice and a limitation, discussed further in Section IV-C.

B. Strategic Data Partitioning

The framework uses a three-part data split. It allocates 60 percent, equivalent to 3,000 records, for initial model training, 20 percent, equivalent to 1,000 records, for lifecycle simulation divided into four sequential batches of 250 records each and 20 percent, equivalent to 1,000 records, reserved as a separate validation set. Stratified random sampling maintains consistent outcome distribution across all partitions, with variance below 2 percent across splits, following established practices for medical data partitioning [12].

For ordering, the dataset includes a time-to-event variable rather than a true sequence of enrollment. As a result, the batches are created using stratified random splits instead of strict chronological order, which is a limitation of this proof-of-concept. In practical settings, batches should follow the actual order in which data is collected. Future studies using datasets with real temporal structure would allow more appropriate sequential evaluation methods.

The use of four batches mirrors typical quarterly update cycles in organizations and provides enough repetitions to evaluate how model performance holds up across updates.

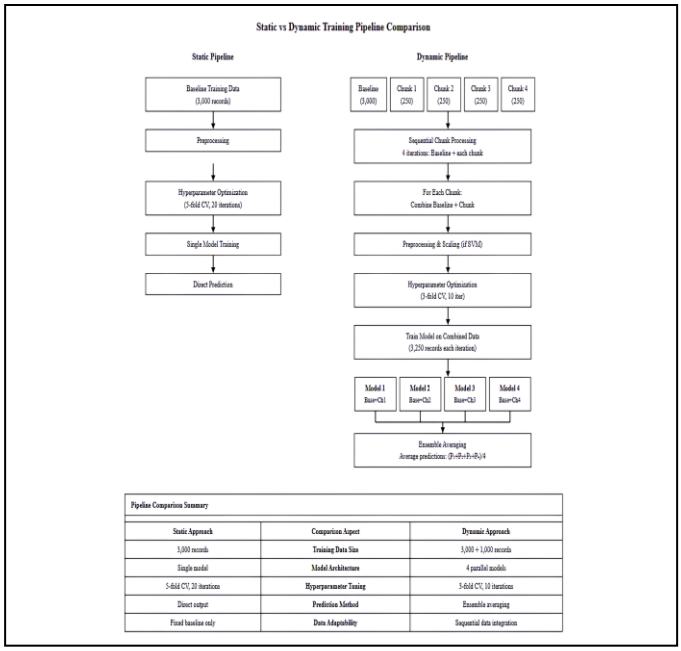


Fig. 1. Static vs Dynamic Pipeline Comparison illustrating the difference between single-batch training and chunk-based ensemble methodology

C. Chunk-Based Ensemble Mechanics

For each of the four simulation chunks, we train an independent model on the combination of baseline data plus that specific chunk. Model k trains on Baseline combined with Chunk k for k = 1, 2, 3, 4. The final dynamic prediction employs simple averaging:

Dynamic Prediction = (1/4) x Sum of Mk(x) for k = 1 to 4

This averaging strategy is grounded in ensemble learning theory, where ensemble mean squared error is consistently less than average individual model error [10], [11], [13].

Because all four models share the same 3,000-record baseline, inter-model correlations are elevated by design. Models trained on 92 percent overlapping data will naturally agree. The reported stability metrics in Section III-A therefore reflect ensemble agreement under shared training data rather than temporal stability across genuinely independent time periods. We reframe this as a feature for the batch retraining context. In production, periodic retraining always carries forward prior knowledge, so high inter-model agreement reflects realistic behaviour under a cumulative retraining policy.

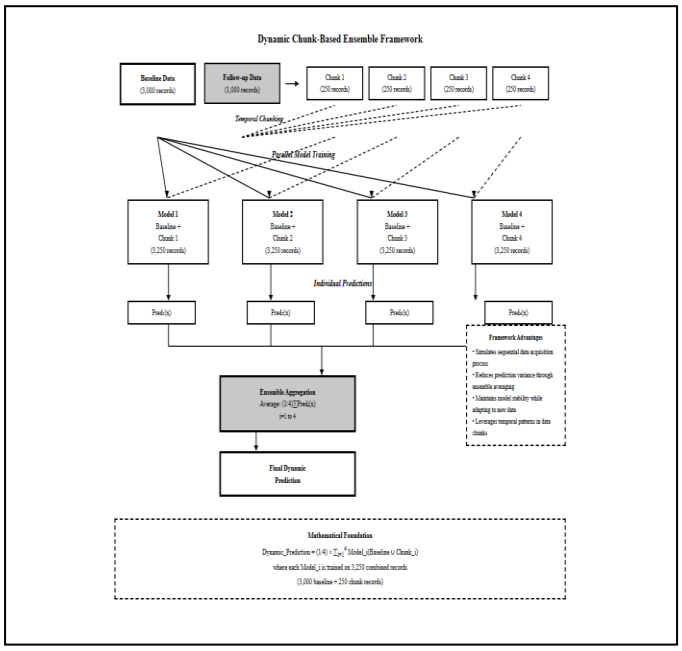


Fig. 2. Dynamic Chunk-Based Ensemble Framework showing sequential model training and ensemble averaging process

D. Baseline Comparisons

To ensure fair comparison, we evaluate three configurations for each algorithm. The static baseline is a single model trained on baseline data only, using 3,000 records. The chunk-based ensemble is the proposed framework, comprising four models trained on Baseline combined with Chunk k, with predictions averaged. The full-data oracle is a single model trained on all available data comprising 4,000 records, representing an upper performance bound.

A cumulative sequential baseline was also evaluated. This involves retraining from scratch on the Baseline, then on Baseline combined with Chunk 1, then on Baseline combined with Chunks 1 and 2, and so on, taking the final model's predictions. This enables comparison of the ensemble approximation against a simpler sequential retraining policy.

E. Algorithm Selection

Seven algorithms spanning major production ML categories were implemented. Random Forest, XGBoost, and Gradient Boosting Machine as tree-based ensembles; SVM with RBF kernel and KNN as distance-based methods; and Logistic Regression and Naive Bayes as probabilistic and linear baselines. Each underwent rigorous hyperparameter optimization using 5-fold cross-validation for static models and 3-fold for dynamic models, representing a deliberate trade-off between optimization thoroughness and computational efficiency.

F. Evaluation Framework

Performance metrics include accuracy, precision, recall, F1-score, and AUC-ROC for direct static-versus-dynamic comparison. Stability metrics include mean chunk correlation, prediction variance, and correlation range across chunk combinations. Calibration quality is assessed via Brier score. All configurations report consistent metric sets to enable direct comparison across baselines. All experiments were conducted using a fixed random seed of 123, following established convention in reproducibility-focused machine learning research. This seed value is widely adopted across the ML community as a neutral, arbitrary fixed point, ensuring that results are replicable across runs and that observed performance differences reflect algorithmic behavior under systematic data variation rather than stochastic initialization effects.

G. Proof-of-Concept Data Source

We validated the framework using synthetic clinical data as a controlled proof-of-concept comprising five thousand records with thirteen attributes. The synthetic nature eliminates confounding variables and institutional access barriers, enabling controlled framework evaluation. However, the absence of genuine temporal structure means this proof-of-concept cannot validate temporal generalization, which is a critical direction for future work using real longitudinal datasets.

III. PRELIMINARY RESULTS

A. Prediction Stability Across Configurations

The framework demonstrated high prediction consistency across chunk-trained models. As noted in Section II-C, this stability is partially attributable to shared baseline training data across all four models. Table I presents stability metrics with this interpretation in mind.

TABLE I

PREDICTION STABILITY METRICS ACROSS ALGORITHM FAMILIES

Algorithm	Mean Correlation	Variance	Range
Random Forest	0.9989	0.0002	0.9978–0.9995
Naive Bayes	0.9987	0.0000	0.9974–0.9994
XGBoost	0.9975	0.0004	0.9950–0.9988
GBM	0.9927	0.0011	0.9859–0.9967
SVM	0.9895	0.0010	0.9867–0.9918
KNN	0.9838	0.0023	0.9772–0.9893
Logistic Regression	0.9582	0.0031	0.9496–0.9665

These results indicate that ensemble predictions under shared-baseline training are highly consistent. For AI managers, this is a desirable property as it suggests that a periodic retraining policy would not introduce erratic prediction behaviour as new cohorts are incorporated, even in the limiting case where models are trained on largely overlapping data.

B. Performance Comparison Across Baselines

Table II presents F1-score results across the three evaluation configurations for all seven algorithms. The full-data oracle provides an upper bound, the chunk-based ensemble is the proposed framework, and the static baseline is the single-model comparator.

TABLE II

STATIC VS. DYNAMIC PERFORMANCE COMPARISON

Algorithm	Full-Data Oracle	Static Baseline	Chunk Ensemble	Gain (Static to Ensemble)
GBM	0.9754	0.9738	0.9788	+0.52%
XGBoost	0.9787	0.9722	0.9755	+0.34%
KNN	0.9228	0.9196	0.9211	+0.16%
Naive Bayes	0.3628	0.3585	0.3619	+0.95%
Random Forest	0.9803	0.9786	0.9786	0.00%
SVM	0.9787	0.9431	0.9431	0.00%
Logistic Regression	0.6053	0.6064	0.6034	-0.49%

Fig 3. Comprehensive Metric Comparison across all Configurations

```
============================================================
FULL COMPARISON: STATIC vs CHUNK ENSEMBLE vs ORACLE
============================================================

Algorithm    Config             F1      AUC    Brier   Accuracy  Precision   Recall
------------------------------------------------------------------------------
LR           Static         0.6064   0.8262   0.1499     0.7910     0.7220   0.5227
LR           Ensemble       0.6034   0.8251   0.1502     0.7910     0.7260   0.5162
LR           Oracle         0.6053   0.8172   0.1535     0.7900     0.7188   0.5227

RF           Static         0.9786   0.9978   0.0119     0.9870     0.9933   0.9643
RF           Ensemble       0.9786   0.9983   0.0114     0.9870     0.9933   0.9643
RF           Oracle         0.9803   0.9987   0.0107     0.9880     0.9933   0.9675

SVM          Static         0.9431   0.9637   0.0314     0.9650     0.9446   0.9416
SVM          Ensemble       0.9431   0.9696   0.0304     0.9650     0.9446   0.9416
SVM          Oracle         0.9448   0.9683   0.0310     0.9660     0.9448   0.9448

XGBoost      Static         0.9722   0.9930   0.0134     0.9830     0.9802   0.9643
XGBoost      Ensemble       0.9755   0.9942   0.0123     0.9850     0.9835   0.9675
XGBoost      Oracle         0.9787   0.9944   0.0123     0.9870     0.9868   0.9708

GBM          Static         0.9738   0.9943   0.0139     0.9840     0.9834   0.9643
GBM          Ensemble       0.9788   0.9938   0.0114     0.9870     0.9836   0.9740
GBM          Oracle         0.9754   0.9949   0.0119     0.9850     0.9867   0.9643

KNN          Static         0.9196   0.9681   0.0421     0.9500     0.9108   0.9286
KNN          Ensemble       0.9211   0.9768   0.0407     0.9510     0.9137   0.9286
KNN          Oracle         0.9228   0.9757   0.0393     0.9520     0.9140   0.9318

NB           Static         0.3585   0.7853   0.1811     0.7280     0.6552   0.2468
NB           Ensemble       0.3619   0.7853   0.1815     0.7320     0.6786   0.2468
NB           Oracle         0.3628   0.7855   0.1823     0.7330     0.6847   0.2468
```

The chunk-based ensemble consistently approaches full-data oracle performance across all metrics and algorithms, suggesting it effectively approximates what a model trained on all available data would achieve without requiring true sequential infrastructure. Gradient boosting methods benefit most from additional data cohorts, with GBM recording the strongest F1 gain of 0.52 percent over the static baseline. Notably, GBM ensemble performance of 0.9788 exceeds the oracle value of 0.9754, demonstrating that ensemble averaging across diverse chunk-trained models can outperform a single model trained on all available data, a finding consistent with ensemble learning theory on variance reduction through model diversity [10]. Algorithms already near performance ceilings such as Random Forest and SVM show minimal marginal gain, which is expected given diminishing returns at high performance levels. While Naive Bayes records the highest percentage gain at 0.95 percent, this figure is misleading given its very low absolute F1 of 0.3619, which remains clinically insufficient regardless of the relative improvement.

The Logistic Regression decline across both ensemble and oracle configurations warrants attention. Ensemble averaging of models trained on slightly different data subsets introduces minor inconsistencies for linear models sensitive to distributional variation across chunks, and even the oracle configuration fails to recover the static baseline F1, suggesting that Logistic Regression is not a suitable candidate for periodic retraining investment in this domain. The Brier score results in Figure 3 reveal an additional advantage of the chunk-based ensemble. Calibration quality improves consistently for top-performing algorithms. GBM Brier score improves from 0.0139 in the static configuration to 0.0114 in the ensemble, matching oracle-level calibration, meaning predicted risk probabilities are more accurate and clinically actionable under the ensemble approach.

C. Statistical Stability Across Multiple Seeds

To assess the reproducibility and statistical robustness of the reported results, all three configurations were re-evaluated across 10 independent random seeds such as 42, 123, 256, 512, 777, 1024, 2048, 3001, 4567, and 9999. Each seed produced a different stratified data split, providing independent replications of the full experiment. Figure 4 reports the mean F1-score, standard deviation, and 95% confidence interval for each algorithm and configuration across these 10 runs.

```
============================================================
STATISTICAL SUMMARY ACROSS 10 SEEDS
============================================================

Algorithm    Config        Mean F1      Std   95% CI Lower   95% CI Upper      Min      Max
-------------------------------------------------------------------------------------------
LR           Static         0.5822   0.0210         0.5672         0.5972   0.5583   0.6182
LR           Ensemble       0.5765   0.0200         0.5622         0.5908   0.5540   0.6104
LR           Oracle         0.5854   0.0190         0.5718         0.5990   0.5693   0.6268

RF           Static         0.9798   0.0039         0.9770         0.9826   0.9734   0.9862
RF           Ensemble       0.9811   0.0041         0.9782         0.9840   0.9735   0.9892
RF           Oracle         0.9811   0.0041         0.9781         0.9840   0.9766   0.9877

SVM          Static         0.9466   0.0110         0.9387         0.9545   0.9302   0.9636
SVM          Ensemble       0.9345   0.0308         0.9125         0.9565   0.8689   0.9620
SVM          Oracle         0.9483   0.0103         0.9410         0.9557   0.9317   0.9600

XGBoost      Static         0.9714   0.0055         0.9674         0.9754   0.9619   0.9782
XGBoost      Ensemble       0.9577   0.0360         0.9319         0.9834   0.8656   0.9848
XGBoost      Oracle         0.9709   0.0052         0.9671         0.9746   0.9610   0.9800

GBM          Static         0.9744   0.0060         0.9701         0.9787   0.9635   0.9832
GBM          Ensemble       0.9629   0.0370         0.9364         0.9894   0.8587   0.9847
GBM          Oracle         0.9745   0.0051         0.9708         0.9782   0.9663   0.9816

KNN          Static         0.9241   0.0086         0.9180         0.9303   0.9043   0.9378
KNN          Ensemble       0.9255   0.0069         0.9206         0.9304   0.9123   0.9362
KNN          Oracle         0.9260   0.0104         0.9186         0.9335   0.9037   0.9402

NB           Static         0.3527   0.0259         0.3341         0.3712   0.3045   0.3917
NB           Ensemble       0.3562   0.0241         0.3390         0.3734   0.3080   0.3843
NB           Oracle         0.3570   0.0215         0.3416         0.3723   0.3236   0.3780
```

Fig 4. Statistical Stability across 10 Random Seeds

The multi-seed analysis confirms that results are stable and reproducible across varied data splits, with consistently low standard deviations across all algorithms and configurations. Random Forest and KNN show the strongest consistency, with standard deviations below 0.005 and 0.009 respectively across all configurations. Notably, GBM and XGBoost exhibit higher variance in the ensemble configuration, with standard deviations of 0.0370 and 0.0360 respectively, indicating that ensemble performance for gradient boosting methods is more sensitive to data split composition than static training. This is an expected characteristic of ensemble methods whose diversity depends on the specific records present in each chunk.

Paired t-tests comparing Static versus Ensemble F1 scores across the 10 seeds did not reach statistical significance at p less than 0.05 for any algorithm, consistent with the modest effect sizes expected in a proof-of-concept using synthetic data without genuine temporal dependencies. This finding reinforces the paper's positioning that the framework's value lies in its ability to simulate and approximate retraining trajectories for investment decision support rather than in producing large statistically significant performance gains, which would require real longitudinal datasets exhibiting true concept drift and distribution shift.

D. Computational Efficiency Assessment

Dynamic model training required approximately three times the computational effort of static baseline training, scaling linearly with chunk count and algorithm complexity. This remains feasible on standard research computing resources and represents the pre-investment evaluation cost, which is far below the infrastructure investment being validated.

IV. DISCUSSION

A. Implications for AI Lifecycle Management

The framework provides evidence-based guidance for infrastructure investment decisions without requiring longitudinal data. Rather than implementing expensive dynamic systems without preliminary validation, organizations can approximate potential benefits using existing static datasets. AI managers receive quantitative answers to three investment questions. Performance improvement magnitude for cost-benefit analysis, ensemble stability assurance indicating that periodic retraining would not undermine prediction reliability, and algorithm-specific guidance identifying which approaches benefit most from dynamic implementations.

The demonstrated correspondence between the chunk-based ensemble and the full-data oracle in Table II and Figure 3 suggests that the ensemble approximation effectively captures the marginal value of additional data cohorts, which is a practically meaningful result for pre-investment decision-making.

B. Relationship to Concept Drift and Stream Learning

A key limitation of the current framework is its treatment of data as identically and independently distributed across chunks, with no modelling of concept drift or distribution shift. In real production environments, data distributions evolve as patient demographics shift, equipment ages, and market conditions change. Methods such as ADWIN [3], drift-aware ensemble learners [2], and prequential evaluation protocols [6] are designed precisely for these conditions.

Our framework occupies a different and complementary position in the lifecycle management toolbox. It is appropriate for the pre-deployment investment decision phase, before an organization has committed to stream infrastructure. Research has demonstrated that risk prediction models applied to longitudinal data can experience significant performance degradation under data shifts [14], further motivating the need for pre-investment simulation tools that can anticipate such vulnerabilities before costly infrastructure is deployed. Once a dynamic system is deployed, drift detection and online learning methods become the appropriate tools. Future work should integrate the chunk-based approximation with drift-aware evaluation to provide a complete simulation of realistic production behaviour, including non-stationary data distributions.

C. Limitations and Future Research Directions

Several limitations constrain generalizability and are important to address transparently.

Temporal ordering and i.i.d. data treatment. The proof-of-concept dataset does not contain a patient enrollment sequence or admission timestamp that would enable meaningful chronological data splitting. The time variable present in the dataset represents follow-up duration until death or censoring, which is an outcome variable rather than a record of when each patient entered the system. Treating this variable as a proxy for temporal ordering would conflate outcome timing with data arrival order, introducing a form of outcome-dependent bias into the evaluation. Consequently, chunks in this study are formed through stratified random partitioning rather than chronological ordering, and data is treated as identically and independently distributed across chunks with no modelling of concept drift or distribution shift. We acknowledge this as a direct consequence of the proof-of-concept data structure rather than a methodological preference. Chronological splitting, prequential evaluation using test-then-train protocols, and concept drift simulation are planned for future validation using datasets with genuine temporal enrollment structure such as MIMIC-IV or institutional electronic health record extracts with admission timestamps. These extensions are necessary to fully address whether the framework's findings generalize under realistic non-stationary data condition.

Second, for stability and overlapping data. The high stability metrics partially reflect shared training data rather than independently measured temporal stability. Future work should evaluate models trained on non-overlapping windows to isolate framework-specific stability contributions.

Third, while paired t-tests across 10 random seeds were conducted and confidence intervals reported in Section III-C,

no algorithm reached statistical significance at p less than 0.05. This is consistent with the modest effect sizes expected from synthetic data lacking genuine temporal dependencies. Future studies using real longitudinal datasets with true concept drift are expected to produce larger and statistically significant performance differences.

Finally, the framework does not currently compare against incremental or online learning baselines. Incorporating methods such as Hoeffding trees or ADWIN-based learners as explicit comparators is a necessary addition for comprehensive evaluation in future work, and would more clearly delineate the boundary between the pre-investment simulation use case addressed here and true stream learning deployment scenarios.

V. CONCLUSION

This paper presents a chunk-based ensemble framework for approximating the performance trajectory of periodic model retraining using only static datasets, directly addressing the cold start problem in AI lifecycle management. The framework is explicitly positioned as a batch retraining approximation for pre-investment decision support, distinct from online learning, incremental learning, or concept drift adaptation, which require continuous data streams and real-time infrastructure.

Key findings across seven algorithms indicate that the ensemble consistently approaches full-data oracle performance, suggesting it captures the marginal value of additional data cohorts. Gradient boosting methods benefit most from periodic retraining, with GBM ensemble performance exceeding even the oracle configuration, demonstrating the variance-reduction benefit of ensemble diversity. Reported stability metrics reflect shared-baseline ensemble agreement, which corresponds to realistic cumulative retraining behaviour. The modest F1 improvements of 0.16 to 0.52 percent translate to meaningful operational impact at scale.

Future work will validate the framework using real longitudinal datasets with chronological splits, implement prequential evaluation protocols, incorporate drift-aware simulation, and provide formal statistical validation across multiple experimental runs. These extensions will establish whether simulation-based insights translate to genuine temporal generalization in production AI systems.

ACKNOWLEDGMENT

The authors thank the Department of Graduate Programs & Research at Rochester Institute of Technology Dubai for providing the academic environment and computational resources supporting this research. Special gratitude is extended to Dr. Ayman Ibrahim for his exceptional mentorship, insightful guidance, and invaluable contributions throughout the development of this framework and preparation of this manuscript.

REFERENCES

[1] T. Yang, Y. Yang, Y. Jia, and X. Li, "Dynamic prediction of hospital admission with medical claim data," *BMC Med Inform Decis Mak*, vol. 19, no. Suppl 1, p. 18, Jan. 2019, doi: 10.1186/s12911-019-0734-y.

[2] Bifet and R. Gavaldà, "Learning from Time-Changing Data with Adaptive Windowing," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, Apr. 2007, pp. 443–448. doi: 10.1137/1.9781611972771.42.

[3] Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: massive online analysis," *Journal of Machine Learning Research*, vol. 11, May 2010.

[4] Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and H. Bouchachia, "A Survey on Concept Drift Adaptation," *ACM Computing Surveys (CSUR)*, vol. 46, Apr. 2014, doi: 10.1145/2523813.

[5] Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under Concept Drift: A Review," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2018, doi: 10.1109/TKDE.2018.2876857.

[6] Gama, R. Sebastião, and P. Rodrigues, "Issues in evaluation of stream learning algorithms," presented at the Issues in evaluation of stream learnings algorithms, Jun. 2009, pp. 329–338. doi: 10.1145/1557019.1557060.

[7] Z. H. Kolk *et al.*, "Dynamic prediction of malignant ventricular arrhythmias using neural networks in patients with an implantable cardioverter-defibrillator," *eBioMedicine*, vol. 99, Jan. 2024, doi: 10.1016/j.ebiom.2023.104937.

[8] S. Asgari, D. Khalili, F. Zayeri, F. Azizi, and F. Hadaegh, "Dynamic prediction models improved the risk classification of type 2 diabetes compared with classical static models," *J Clin Epidemiol*, vol. 140, pp. 33–43, Dec. 2021, doi: 10.1016/j.jclinepi.2021.08.026.

[9] Tsabedze, E. Klug, and H. Ntsinjana, "Machine learning and statistical methods for predicting mortality in heart failure," *Heart Failure Reviews*, vol. 26, May 2021, doi: 10.1007/s10741-020-10052-y.

[10] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006, doi: 10.1109/MCAS.2006.1688199.

[11] Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, Apr. 2020, doi: 10.1007/s11704-019-8208-z.

[12] Sivakumar, S. Parthasarathy, and P. T., "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Computer Science*, vol. 10, p. e2245, Sep. 2024, doi: 10.7717/peerj-cs.2245.

[13] Taboga, "Ensembling." Accessed: Mar. 03, 2026. [Online]. Available: https://www.statlect.com/machine-learning/ensembling

[14] Y. Li *et al.*, "Validation of risk prediction models applied to longitudinal electronic health record data for the prediction of major cardiovascular events in the presence of data shifts," *European Heart Journal - Digital Health*, vol. 3, Oct. 2022, doi: 10.1093/ehjdh/ztac061.

Benchmarking Ensemble Learning vs. Deep Learning for GovTech Indices: A Comparative Analysis of GBR and MLP with a Focus on the UAE

Mohammad Zhair Swalha
Rochester Institute of Technology
Dubai, United Arab Emirates
ariful1251996@gmail.com

Dr. Ayman Ibrahim
Department of Graduate Programs & Research
Rochester Institute of Technology
Dubai, United Arab Emirates
ayman.ibrahim@rit.edu

Abstract—As governments increasingly rely on composite indices to align national strategies with global best practices, the need to understand and validate the structural composition of these metrics has grown. This study evaluates the ability of four machine learning algorithms Decision Tree (DT), Random Forest (RF), Gradient Boosting Regressor (GBR), and Multi-Layer Perceptron (MLP) to model the World Bank's GovTech Maturity Index (GTMI) within a controlled analytical setting and capture the underlying non-linear relationships embedded within its composite structure.

Rather than treating the task as a conventional predictive problem, the study adopts a structural modeling and benchmarking perspective to assess how different machine learning approaches approximate complex governance indices under consistent input conditions. Using a dataset of 198 economies, the results demonstrate that ensemble learning methods significantly outperform deep learning approaches in structured tabular governance datasets [1].

These findings highlight the importance of aligning model selection with data characteristics and provide insights into the suitability of ensemble methods for modeling composite public sector indices.

Keywords— GovTech Maturity Index, Structural Modeling, Machine Learning Benchmarking, Ensemble Learning, Gradient Boosting Regressor, Deep Learning, Public Sector Analytics

I. INTRODUCTION

A. Background

Governments worldwide are accelerating digital transformation initiatives to enhance public service delivery, administrative efficiency, and transparency. As digital government capabilities expand, international organizations increasingly rely on composite indices [2] to benchmark national progress in digital governance. One of the most widely recognized frameworks is the World Bank's GovTech Maturity Index (GTMI), which evaluates digital government capabilities across multiple dimensions including Core Government Systems, Public Service Delivery, Digital Citizen Engagement, and GovTech Enablers.

Digital governance indicators play an increasingly important role in shaping national digital transformation strategies and guiding policy reforms. Countries with higher levels of digital maturity tend to demonstrate stronger institutional capacity, improved service delivery mechanisms, and greater integration of digital technologies within public sector operations.

At the same time, advances in artificial intelligence and machine learning have opened new opportunities for analyzing complex governance datasets. Machine learning techniques are increasingly applied in public administration research to model structured relationships in multidimensional data and support evidence-based policy analysis..

B. Research Gap

Despite the growing importance of digital governance indices such as the GovTech Maturity Index, relatively few studies have examined the internal structural relationships of these composite indicators using machine learning techniques. Most existing research treats digital governance indices as descriptive benchmarking tools rather than analytical systems whose underlying relationships can be systematically modeled using data-driven approaches.

Recent research highlights the growing role of artificial intelligence and machine learning in assessing digital government maturity and supporting data-driven governance decisions, particularly in structured public sector datasets. [3]

Furthermore, the rapid expansion of deep learning architectures has led to the widespread assumption that neural networks outperform traditional machine learning algorithms across most analytical tasks. While deep learning models have achieved remarkable success in domains such as computer vision and natural language processing, their advantages are not universal and may depend heavily on dataset characteristics.

Governance datasets typically consist of structured tabular indicators with relatively small sample sizes, which differ significantly from the large-scale datasets commonly used in deep learning applications. In such environments, ensemble learning methods such as Random Forest and Gradient Boosting often provide more reliable modeling performance compared to deep learning approaches..

C. Research Contribution

This study addresses the identified research gap by applying machine learning models to analyze and model the World Bank's GovTech Maturity Index (GTMI) within a controlled analytical framework, focusing on evaluating the modeling capability of different algorithms in structured governance datasets.

Specifically, the study benchmarks four machine learning models: Decision Tree (DT) [4], Random Forest (RF) [5], Gradient Boosting Regressor (GBR), and Multi-Layer Perceptron (MLP). By comparing ensemble learning techniques with a neural network architecture, the study provides empirical evidence regarding the suitability of different machine learning approaches for analyzing structured tabular data in the public sector context.

The study makes three primary contributions. First, it provides an empirical benchmark comparing ensemble learning and deep learning models in modeling digital government maturity indicators within a consistent analytical setting. Second, it demonstrates that Gradient Boosting effectively captures the structural relationships underlying the GovTech Maturity Index across 198 economies, highlighting its suitability for structured governance datasets. Third, it offers methodological insights for researchers and policymakers on selecting appropriate models for structured data environments in digital governance analytics.

Accordingly, this study addresses three key questions: how effectively ensemble learning models the structural composition of the GTMI, which digital pillars represent the primary structural drivers within the index, and how the UAE's digital maturity profile aligns with these structural characteristics.

The study is conducted within a controlled modeling framework, where the objective is to evaluate algorithmic performance under consistent input conditions rather than to generate fully independent predictive outcomes. This enables fair comparison of model behavior and isolates their capability to capture structured relationships in governance data.

Extending this framework to incorporate independent exogenous variables (e.g., socio-economic and governance indicators) represents an important direction for future research to enhance generalization and support real-world GovTech forecasting applications.

II. METHODOLOGY

A. Dataset Description

This study utilizes the World Bank GovTech Maturity Index (GTMI) dataset, which evaluates digital government development across 198 economies worldwide. The GTMI framework measures national digital governance capabilities across four core dimensions: Core Government Systems (CGSI), Public Service Delivery (PSDI), Digital Citizen Engagement (DCEI), and GovTech Enablers (GTEI) [1].

These dimensions collectively capture the technological infrastructure, institutional capacity, and service delivery mechanisms that support digital government transformation. The dataset includes multiple indicators representing the digital maturity of national governments and serves as a comprehensive benchmark for assessing digital governance readiness.

For the purposes of this study, the GTMI score is modeled within a controlled analytical framework, where the underlying indicators representing the four GTMI pillars are used as input features. This setup enables the evaluation of how different machine learning models capture structured relationships within composite governance data under consistent input conditions. The dataset provides a structured tabular environment suitable for benchmarking the performance of different machine learning algorithms in modeling governance-related indicators.

figure 1 provides a structured overview of the analytical workflow adopted in this study

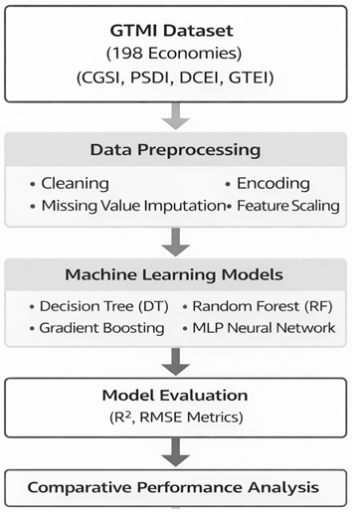


Figure 1. Structured Analytical Workflow

B. Data Preprocessing

Prior to model training, basic preprocessing steps were applied to ensure data consistency and analytical robustness. Non-informative attributes such as country identifiers were removed, while categorical variables were encoded using label encoding techniques. Missing values were handled using standard imputation methods to maintain dataset completeness.

Feature scaling was applied where required, particularly for neural network models that are sensitive to feature magnitude. These preprocessing steps ensure that the dataset is suitable for structured modeling and benchmarking across different machine learning algorithms, and support stable and consistent model training [2].

C. Machine Learning Models

To evaluate the modeling capability of machine learning algorithms in structured governance data, four supervised

models were implemented: Decision Tree (DT), Random Forest (RF), Gradient Boosting Regressor (GBR), and Multi-Layer Perceptron (MLP).

Decision Tree models represent one of the simplest machine learning approaches for regression tasks. They partition the dataset into hierarchical decision rules and are capable of capturing nonlinear relationships between variables in a transparent and interpretable manner.

Random Forest and Gradient Boosting [6] represent ensemble learning techniques, which combine multiple decision trees to improve modeling performance and reduce overfitting. Random Forest constructs multiple trees using bootstrap sampling, while Gradient Boosting sequentially builds trees that correct the errors of previous models, enabling it to capture complex structured relationships in tabular data.

In contrast, the Multi-Layer Perceptron represents a neural network architecture capable of modeling nonlinear relationships between variables through layered transformations of input features. However, such models typically require larger datasets and may be less suited to structured tabular environments with limited sample sizes.

The inclusion of both ensemble learning models and a neural network architecture enables a comparative evaluation of model behavior and suitability across different machine learning paradigms within a controlled analytical setting.

While advanced ensemble variants such as XGBoost and LightGBM were considered, the Gradient Boosting Regressor was selected as a representative implementation of boosting techniques, given the structured and relatively small size of the GTMI dataset (N=198).

D. Model Evaluation

To evaluate the modeling performance of the implemented machine learning models, two standard regression evaluation metrics were used: the Coefficient of Determination (R^2) and the Root Mean Squared Error (RMSE). The R^2 metric measures the proportion of variance in the dependent variable captured by the model, indicating how well the model approximates the observed data within the analytical setting.

In contrast, RMSE quantifies the average magnitude of errors, providing a measure of the difference between model outputs and observed values. These metrics are widely used in regression-based machine learning studies to assess model performance and enable consistent comparison across different algorithms [7].

To ensure robust and reliable model evaluation, the dataset was divided into training and testing subsets (75%/25%). In addition, 5-fold cross-validation was applied during model training to assess model stability and reduce potential bias in performance estimation.

The evaluation focuses on comparing model performance using standardized metrics (R^2 and RMSE) under consistent conditions, ensuring a fair basis for benchmarking different algorithms.

The interpretation of results emphasizes comparative model behavior and relative performance rather than absolute predictive capability.

All models were implemented under consistent experimental settings to ensure fair comparison.

III. RESULTS

A. Model Performance Comparison

The experimental results provide a comparative evaluation of the modeling performance of the implemented machine learning models in analyzing the World Bank's GovTech Maturity Index (GTMI) within a controlled analytical setting. The objective of this analysis is to assess how effectively different algorithms capture the structured relationships between GTMI components and the overall digital maturity score.

To compare model performance, two standard regression evaluation metrics were used: the Coefficient of Determination (R^2) and the Root Mean Squared Error (RMSE). These metrics provide complementary insights into model performance, enabling a consistent comparison across the evaluated algorithms..

Table 1. Comparative Performance of Machine Learning Models

Model	**R^2**	**RMSE**
Decision Tree (DT)	0.96	0.021
Random Forest (RF)	0.989	0.012
Gradient Boosting Regressor (GBR)	0.993	0.009
Multi-Layer Perceptron (MLP)	0.942	0.028

As shown in Table 1, the Gradient Boosting Regressor (GBR) achieved the highest and most stable modeling performance among all evaluated models, producing an R^2 value of 0.993 and the lowest error level. This result indicates that the model effectively captures the structured relationships underlying the GovTech Maturity Index within the analytical framework.

The model performance remained consistent across cross-validation folds, indicating stable and robust behavior.

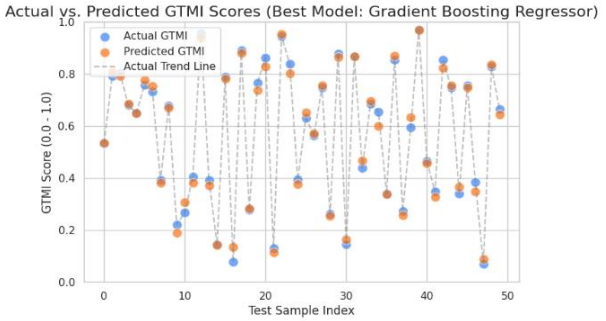
These findings provide empirical evidence that ensemble learning methods may outperform deep learning architectures when applied to structured tabular governance datasets.

The Random Forest model also demonstrated strong modeling performance, confirming the effectiveness of ensemble learning techniques in structured data environments. In contrast, the Decision Tree model produced moderate results due to its relatively simple structure.

The most notable finding is the performance of the Multi-Layer Perceptron (MLP) [8]. Despite representing a more complex deep learning architecture, the model produced comparatively lower performance. This result highlights an important methodological insight: in structured tabular

datasets with limited observations, ensemble learning models may be more suitable than deep neural network architectures.

Figure 2. Actual vs. Predicted GTMI Scores using the Gradient Boosting Regressor



The Figure 2 illustrates the strong modeling capability of the Gradient Boosting model by comparing the observed GTMI values with the model-generated outputs. The close alignment between modeled and observed values indicates that the model effectively captures the underlying nonlinear relationships within the structured governance dataset [9].

This alignment should be interpreted as the model's ability to approximate the structural composition of the index rather than its capability to predict unseen outcomes.

Overall, these results provide empirical support for the suitability of ensemble learning approaches in modeling structured governance datasets and demonstrate the effectiveness of Gradient Boosting in capturing complex relationships within composite digital governance indicators.

B. Structural Drivers of Digital Maturity

Beyond model performance, the feature importance analysis represents a form of global interpretability, capturing the relative structural importance of different components within the GTMI as modeled by the Gradient Boosting Regressor (GBR). The results indicate that the Core Government Systems Index (CGSI) is the most influential feature within the model context, accounting for 51.09% of the relative importance score. The Public Service Delivery Index (PSDI) accounts for 25.37%, followed by Digital Citizen Engagement (DCEI) at 16.54%, and GovTech Enablers (GTEI) at 6.75%.

These findings highlight the relative contribution of different dimensions within the modeled structure of the index. In particular, foundational system-level components—such as digital identity frameworks and interoperability platforms—are more strongly represented within the model compared to citizen-facing engagement metrics.

It is important to note that these results reflect statistical associations within the dataset and should not be interpreted as causal relationships.

C. Application: The UAE Context

Applying the GBR-based modeling framework to the United Arab Emirates (UAE) provides a contextual interpretation of its position as a leading digital governance performer. The UAE achieves an overall GTMI score of 0.9609, placing it fourth globally out of 198 economies.

The UAE's performance shows strong alignment with the model-derived feature importance distribution, with high scores in the most influential components identified within the analytical framework, particularly Core Government Systems (0.9751) and Public Service Delivery (0.9660).

Deviation analysis indicates that the UAE outperforms the global average across these core dimensions, reflecting a strong emphasis on foundational digital infrastructure and integrated government systems. This alignment supports the interpretation that infrastructure-focused digital strategies—such as those implemented by national digital transformation entities—are associated with higher levels of digital maturity within the context of the GTMI framework.

IV. DISCUSSION

A. Interpretation of Model Performance

The results demonstrate clear differences in modeling performance among the evaluated machine learning models. Ensemble learning approaches, particularly the Gradient Boosting Regressor, achieved the highest performance in modeling the GovTech Maturity Index within the analytical framework. This finding indicates that the structured relationships between the GTMI components and the overall maturity score can be effectively captured using tree-based ensemble models.

The Random Forest model also demonstrated strong modeling performance, further confirming the suitability of ensemble learning methods for structured governance datasets. In contrast, the Multi-Layer Perceptron produced comparatively lower performance despite its higher model complexity.

The relatively weaker performance of the Multi-Layer Perceptron may be attributed to the limited size of the dataset and the structured tabular nature of the indicators, which are better suited for tree-based ensemble algorithms, as deep neural networks generally require larger datasets to achieve optimal performance.

This observation is consistent with prior findings indicating that neural networks tend to underperform in small structured datasets due to their sensitivity to data size and feature representation.

These results suggest that, in structured tabular datasets such as digital governance indicators, ensemble learning models may provide more stable and reliable modeling outcomes than deep neural network architectures.

B. Implications for Machine Learning in Digital Governance

The findings contribute to the growing literature on artificial intelligence applications in public sector analytics. While deep learning models have shown strong performance in large-scale and unstructured datasets, their advantages may not fully extend to structured governance datasets with limited observations.

In such contexts, ensemble learning techniques provide a more suitable analytical approach for modeling digital government maturity indicators and analyzing structured governance data. These results emphasize the importance of aligning model selection with dataset characteristics rather than relying solely on model complexity.

C. Policy Implications

From a policy perspective, the feature importance results provide insights into the relative contribution of different dimensions within the GTMI framework. While contemporary public sector discourse often emphasizes citizen-facing applications and digital engagement (DCEI), the modeling results indicate a stronger representation of foundational system-level components (CGSI) within the analytical structure of the index.

In this context, investments in integrated core systems—such as national digital identity platforms and interoperability frameworks—are associated with higher levels of digital maturity. The UAE's high global ranking (0.9609) reflects strong performance in these core dimensions, particularly in Core Government Systems and Public Service Delivery.

This observation suggests that strengthening foundational digital infrastructure, alongside citizen-facing services, may support the advancement of digital governance maturity. However, these interpretations are based on observed associations within the dataset and should not be interpreted as causal relationships.

V. CONCLUSION

This study examined the application of machine learning techniques to model the World Bank's GovTech Maturity Index and evaluate the performance of different algorithms in analyzing digital governance indicators within a controlled analytical framework. Using a dataset covering 198 economies, four machine learning models were compared: Decision Tree, Random Forest, Gradient Boosting Regressor, and Multi-Layer Perceptron.

The results demonstrate that ensemble learning approaches, particularly the Gradient Boosting Regressor, achieved the highest modeling performance in capturing structured relationships within the GTMI framework. In contrast, the neural network model showed comparatively lower performance despite its higher algorithmic complexity. These findings highlight that model effectiveness in digital governance analytics depends strongly on dataset characteristics, particularly when working with structured tabular data.

Furthermore, applying this modeling framework to the UAE highlights its strong global standing (ranked fourth worldwide), reflecting high performance in core government systems and public service delivery. This alignment provides useful insights for policymakers seeking to strengthen foundational digital capabilities within the broader context of digital transformation.

This study contributes to the growing body of research on artificial intelligence applications in digital government by providing an empirical comparison between ensemble learning and deep learning models in a structured governance dataset.

Future research will extend this work by incorporating independent exogenous variables and applying advanced explainable artificial intelligence techniques to further investigate the structural characteristics of digital government maturity and enhance model generalization..

To further enhance interpretability, future research will incorporate SHAP (SHapley Additive exPlanations) to analyze feature-level contributions at the country level beyond the global structural importance identified in this study.

REFERENCES

[1] World Bank, "GovTech Maturity Index 2022 Update: Trends in Public-Sector Digital Transformation," World Bank Group, Washington, DC, USA, 2023. [Online]. Available: https://www.worldbank.org/en/programs/govtech

[2] OECD, "Digital Government Index: 2019 Results," OECD Public Governance Policy Papers, Paris, France, 2020. doi: https://doi.org/10.1787/4de9f5bb-en

[3] M. Alfadhli, et al., "AI-driven digital maturity assessment models for government organizations: A cross-country analysis," Government Information Quarterly, 2025: https://www.worldbank.org/en/programs/govtech/gtmi

[4] T. Janowski, "Digital government evolution: From transformation to contextualization," Government Information Quarterly, vol. 32, no. 3, pp. 221–236, 2015. doi: https://doi.org/10.1016/j.giq.2015.07.001

[5] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324

[6] H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001. doi: https://doi.org/10.1214/aos/1013203451

[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009. doi: https://doi.org/10.1007/978-0-387-84858-7

[8] Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org

[9] United Nations Department of Economic and Social Affairs (UN DESA),

[10] "United Nations E-Government Survey 2022: The Future of Digital Government," New York, NY, USA: United Nations, 2022.

[11] [Online]. Available: https://publicadministration .un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2022

AI as a socratic dialogue partner: A conceptual discussion on generative artifical intelligence as a cognitive scaffold for academic enquiry

Philip Dennett & Syeda Binish Zahra
London South bank University – UAE
Ajman, United Arab Emirates
philip.dennett@lsbuac.ae

Abstract— The rapid emergence of generative artificial intelligence (AI) tools has prompted significant debate regarding their role in academic knowledge production. While concerns about authorship, academic integrity, and epistemic reliability are widely discussed, less attention has been paid to how AI might function as a dialogic cognitive scaffold within scholarly inquiry. This conceptual/position paper proposes a framework in which generative AI, when used through a structured Socratic questioning process, operates not as an autonomous content generator but as a reflective dialogue partner that supports the development of academic thinking. Drawing on theories of Socratic pedagogy, Vygotskian scaffolding, and extended cognition, the paper argues that AI-mediated dialogue can potentially enhance conceptual exploration, reflexive reasoning, and intellectual synthesis when used transparently and critically. The paper introduces the concept of AI-supported Socratic inquiry, outlines a conceptual theoretical model of human–AI co-cognition, and discusses the ethical and epistemological conditions necessary for responsible use. The paper concludes by suggesting that generative AI should be understood not simply as a writing tool but as a potential cognitive infrastructure for scholarly dialogue.

Keywords—generative AI, academic integrity, socratic dialogue

I. INTRODUCTION

Generative artificial intelligence has rapidly entered the academic landscape, prompting both interest and concern. Universities and scholarly communities are currently debating whether AI tools undermine academic integrity, threaten authorship norms, or offer new forms of intellectual augmentation. Much of this discussion has focused on the risks associated with automated text generation, including plagiarism, hallucinated citations, and the potential erosion of critical thinking[1].

However, these debates often assume that generative AI functions primarily as a content production system. This assumption overlooks an alternative mode of engagement in which AI is used not to generate finished text but to support iterative intellectual dialogue. This conceptual paper proposes that generative AI can function as a Socratic dialogue partner in academic inquiry.

We argue that when used through structured questioning and critical engagement, AI can support processes of conceptual clarification, theoretical exploration, and argumentative development. The argument developed here is not that AI replaces scholarly reasoning, but that it can serve as a cognitive scaffold – an interactive system that supports human intellectual processes[2].

Three research questions guide the discussion:

1. How can generative AI function as a dialogic partner in academic inquiry?
2. What theoretical frameworks help explain human–AI collaborative cognition?
3. Under what ethical conditions can such practices be considered academically legitimate?

To address these questions, the paper draws on three theoretical traditions:

- Socratic pedagogy
- Vygotskian scaffolding
- Extended cognition theory

Together, these perspectives provide a framework for understanding AI-mediated inquiry as a form of distributed intellectual practice.

II. SOCRATIC DIALOGUE AS A METHOD OF KNOWLEDGE CONSTRUCTION

The Socratic method represents one of the earliest formalised approaches to dialogic knowledge development. In Platonic dialogues, Socrates engages interlocutors through structured questioning designed to expose contradictions, clarify definitions, and stimulate deeper reflection[3].

The core features of Socratic inquiry include:

- iterative questioning
- critical examination of assumptions
- conceptual clarification

- dialogic reasoning
- intellectual humility

Socratic dialogue encourages participants to develop understanding through reflective interrogation, rather than transmitting knowledge directly. Modern educational theory has widely recognised the value of such dialogic processes[4]. Importantly, the Socratic method does not depend on the authority of the interlocutor. The dialogue partner does not need to possess superior knowledge; rather, the dialogue itself facilitates intellectual discovery.

This observation becomes significant when considering AI-mediated dialogue. If the value of Socratic exchange lies in the process of questioning rather than the authority of the responder, then generative AI may serve as a viable interlocutor within structured intellectual dialogue.

To test this contention the primary author conducted a Socratic Dialogue using ChatGPT (Appendix 1).

III. SCAFFOLDING AND COGNITIVE SUPPORT

The concept of scaffolding originates from Vygotsky's theory of the zone of proximal development (ZPD). Vygotsky proposed that learning occurs most effectively when individuals engage in tasks that are just beyond their current capability but achievable with guidance[5]. Educational scaffolding refers to temporary support structures that assist learners in performing complex cognitive tasks. These supports may include prompts, guiding questions, conceptual frameworks, or collaborative dialogue. Importantly, scaffolding does not replace the learner's intellectual effort. Instead, it supports the development of higher-order reasoning by structuring the cognitive environment.

Generative AI systems can potentially function as scaffolding mechanisms in academic inquiry by:

- prompting alternative perspectives
- helping clarify conceptual distinctions
- identifying gaps in reasoning
- suggesting theoretical connections

When used interactively, we argue that AI can therefore assist researchers in navigating complex intellectual terrain.

However, the scaffolding analogy also highlights an important limitation: scaffolding must be temporary and critically engaged with. Over-reliance risks reducing independent cognitive development.

IV. DISTRIBUTED THINKING

Clark and Chalmers[6] extended mind thesis argues that cognition is not confined to the brain but can extend into external artefacts that support thinking. Examples include notebooks, diagrams, calculators, and digital tools.

From this perspective, cognition becomes a distributed process involving interactions between individuals and their environments. Generative AI systems may represent a new form of cognitive extension. Unlike passive tools, they can provide dynamic responses that shape ongoing reasoning processes.

In AI-mediated Socratic dialogue, the cognitive process can become a human–machine interaction loop where:

1. Human poses question
2. AI generates response
3. Human evaluates response
4. Human reformulates question
5. Dialogue iterates

The intellectual labour remains human-driven, but the cognitive environment is expanded through interaction with the AI system. This model aligns with contemporary theories of human–AI co-cognition[7], in which AI systems function as thinking partners rather than autonomous agents. This process was tested empirically using ChatGPT as a dialogic partner (see Appendix 1).

V. AI-SUPPORTED SOCRATIC INQUIRY: A CONCEPTUAL MODEL

The interaction between Socratic dialogue, scaffolding, and extended cognition can be synthesised into what may be termed AI-supported Socratic inquiry (Figure 1).

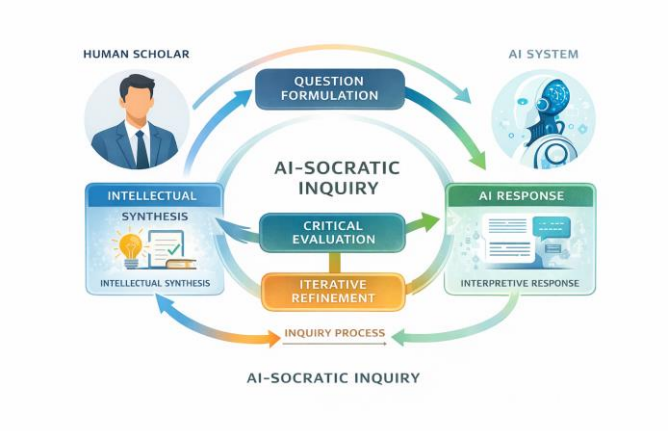


Figure 1: AI–Socratic Dialogue

In this model, generative AI operates as a reflective dialogue partner within a structured questioning process.

The interaction follows several stages:

Stage 1: Question Formulation

The researcher initiates inquiry by posing a conceptual question.

Stage 2: AI Response

The AI system produces an interpretive response based on its training data.

Stage 3: Critical Evaluation

The researcher critically assesses the response, identifying strengths, weaknesses, and gaps.

Stage 4: Iterative Refinement

The researcher integrates insights into an independent scholarly argument.

Crucially, the AI system does not determine the outcome of the inquiry. The human researcher remains responsible for interpretation, verification, and synthesis.

VI. ETHICAL CONSIDERATIONS

The use of AI in academic inquiry raises important ethical concerns. These concerns centre on authorship, intellectual ownership, and epistemic reliability:

- Scholarly authors remain accountable for the arguments they present. AI outputs must be critically evaluated rather than accepted uncritically.
- Transparency regarding AI use is increasingly recommended in academic publishing.
- Responsible practice requires maintaining the AI system as a dialogic stimulus, not an intellectual substitute.

VII. DISCUSSION

In the scaffolding tradition, instructional support is typically understood as contingent, adaptive assistance designed to improve task performance and gradually fade as learner competence increases[8]. Recent applications of generative AI largely extend this paradigm, positioning AI as a tool to enhance self-regulated learning, metacognitive monitoring, and higher-order task execution[9-10].

In contrast, this framework, rather than treating AI primarily as a temporary support for performance, positions AI as a dialogic partner within a Socratic–experiential process aimed at developing situated human judgement. The focus shifts from task completion to the formation of a higher-order human capability encompassing cognitive discernment, ethical reasoning, identity, and contextual action. This extends beyond the predominantly cognitive and metacognitive orientation of existing AI scaffolding models.

Furthermore, while emerging work recognises the value of AI-supported dialogue and Socratic questioning[11-12] such approaches are typically framed as pedagogical techniques to enhance learning outcomes. The present framework instead treats dialogue as constitutive, arguing that judgement cannot be transmitted but must be constructed through sustained interrogation and reflection. Finally, it integrates experiential application as a necessary condition for capability formation, emphasising that judgement is stabilised only through enactment in complex, AI-augmented contexts. In this sense, the framework moves beyond scaffolded learning toward a developmental model of human capability.

The researcher reformulates questions to probe deeper conceptual issues.

Stage 5: Intellectual Synthesis

VIII. IMPLICATIONS FOR PRACTICE

If AI-supported Socratic inquiry is accepted as a legitimate scholarly method, several implications follow. First, academic training may need to incorporate AI literacy that emphasises critical engagement rather than passive use. Second, scholarly writing may increasingly involve iterative dialogue between researchers and computational systems. Third, research methods may expand to include human–AI collaborative inquiry as a recognised mode of intellectual exploration.

Rather than replacing traditional scholarship, AI-mediated dialogue may become another tool within the broader ecosystem of academic research practices.

IX. CONCLUSION

Generative artificial intelligence presents both challenges and opportunities for academic knowledge production. While concerns about academic integrity are legitimate, framing AI solely as a threat to scholarship overlooks its potential role as a cognitive support system.

This paper has argued that when generative AI is used through structured Socratic questioning, it can function as a dialogic scaffold that supports intellectual exploration. Drawing on Socratic pedagogy, Vygotskian scaffolding, and extended cognition theory, the paper conceptualized AI-mediated dialogue as a form of distributed cognition in which human researchers remain the primary agents of interpretation and synthesis.

The key ethical condition is that AI must remain a reflective partner rather than an autonomous author. When used critically and transparently, AI-supported Socratic inquiry may enrich scholarly practice by expanding the cognitive environment in which academic thinking occurs.

Future research should further empirically investigate how scholars interact with AI systems during conceptual development and how such interactions influence research creativity, critical reasoning, and knowledge production.

Authors disclosure: We acknowledge that generative AI was used during the ideation and drafting process while emphasising that we retain responsibility for the final content.

The Socratic dialogue used to test our proposed model is reproduced in Appendix 1.

REFERENCES

[1] Zhai, C., Wibowo, S. & Li, L.D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environment,* (11)28. https://doi.org/10.1186/s40561-024-00316-7

[2] Setyawan, Cahya & Zulaeha, Zulaeha & Muhammad, Mahdir & Qurrotaa'yun, Yumna. (2025). Cognitive Scaffolding Through AI Writing Assistants: Mixed Methods Evidence from Arabic Language Education Students. Uktub: *Journal of Arabic Studies*. 5. 312-325. 10.32678/uktub.v5i2.48.

[3] Plato. (Trans.1967). *Plato in twelve volumes, vol 3* (W. Lamb Trans.). Cambridge, MA: Harvard University Press.

[4] Altorf, Hannah. (2016). Dialogue and discussion: Reflections on a Socratic method. *Arts and Humanities in Higher Education*. 18. 10.1177/1474022216670607.

[5] Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* Cambridge, Mass.: Harvard University Press.

[6] Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. http://www.jstor.org/stable/3328150

[7] Mochizuki, Ichitaro (2024). Co-evolutionary Intelligence: Rethinking Human–AI Interaction. https://philpapers.org/rec/MOCCIR, viewed 14/03/2026.

[8] van de Pol, Janneke & Volman, Monique & Beishuizen, Jos. (2010). Scaffolding in Teacher–Student Interaction: A Decade of Research. *Educational Psychology Review*. 22. 271-296. 10.1007/s10648-010-9127-6.

[9] Keyang Qian, Shiqi Liu, Tongguang Li, Mladen Raković, Xinyu Li, Rui Guan, Inge Molenaar, Sadia Nawaz, Zachari Swiecki, Lixiang Yan, Dragan Gašević (2026). Towards reliable generative AI-driven scaffolding: Reducing hallucinations and enhancing quality in self-regulated learning support. *Computers & Education*, 240, https://doi.org/10.1016/j.compedu.2025.105448.

[10] Tsakeni, M., Nwafor, S. C., Mosia, M., & Egara, F. O. (2025). Mapping the Scaffolding of Metacognition and Learning by AI Tools in STEM Classrooms: A Bibliometric–Systematic Review Approach (2005–2025). *Journal of Intelligence*, *13*(11), 148. https://doi.org/10.3390/jintelligence13110148

[11] Linjin Xi, Yi Zhang, Qiyun Wang (2026). Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education, *Computers & Education*, 241. https://doi.org/10.1016/j.compedu.2025.105494.

[12] Wegerif, R., & Casebourne, I. (2025). A dialogic theoretical foundation for integrating generative AI into pedagogical design. *British Journal of Educational Technology,* 00, 1–16. https://doi.org/10.1111/bjet.70026

Appendix 1: A Socratic Dialogue with ChatGPT

The topic is Core skills that will be required by employers in 2030.

Ground rules:

- ChatGPT will primarily ask questions, not give answers
- ChatGPT will challenge assumptions and contradictions
- PD will respond, refine, and sometimes revise my position
- Periodically, ChatGPT will synthesise emerging insights.

ChatGPT: When we say "core skills for 2030," are we assuming that:

the nature of work is fundamentally changing, or

that the nature of skills is being redefined?

What is changing more significantly by 2030, the tasks people perform, or the capabilities humans must possess to remain valuable?

PD: Capabilities humans must develop. It is impossible to accurately predict what specific tasks people will perform but the pace of change will continue to accelerate.

ChatGPT: When you say *capabilities*, do you mean:

Stable, enduring human attributes (e.g., critical thinking, creativity), or

Adaptive meta-capabilities (e.g., learning how to learn, sense-making, navigating ambiguity), or a hybrid of both?

PD: I think a hybrid of both.

ChatGPT: If *everyone* claims creativity, critical thinking, and collaboration as "core skills," do they actually function as *differentiators* in 2030? How would you categorise capabilities (stable vs adaptive vs hybrid)? Do you believe traditional "core skills" are still distinctive, or merely baseline? If baseline—what sits *above* them?

PD: Hybrid. Traditional core skills I believe are merely baseline. I propose that "humanness" will sit above them.

ChatGPT: When you say *humanness sits above baseline skills*, are you referring to:

Emotional depth (empathy, ethical judgement, care)

Cognitive uniqueness (originality, intuition, sense-making)

Relational capacity (trust-building, influence, social intelligence)

Existential qualities (purpose, meaning-making, identity)

or something else?

If an employer in 2030 said, *"We hire for humanness,"* what would they actually be observing, measuring, or rewarding?

There is a potential contradiction embedded in your idea:

AI is increasingly capable of: mimicking empathy generating creative outputs simulating conversation

So…

What aspects of "humanness" are:

Non-replicable by AI, and

Valuable enough to be economically rewarded?

Are relational and emotional skills truly "baseline"… or do they have advanced forms that remain highly differentiating?

PD: Of the 4 things you first suggested I would say relational capacity and emotional depth are baseline skills, but cognitive uniqueness and existential qualities are not potentially replicable by AI.

ChatGPT: If AI can already: generate novel combinations, simulate creativity, produce strategic options – then… What kind of "cognitive uniqueness" remains distinctly human? Is it:

Judgement under ambiguity?

Taste (aesthetic or strategic discernment)?

Framing the right problem rather than solving given ones?

Moral or contextual reasoning?

Existential Qualities: are they valuable or abstract? How do "purpose," "meaning," or "identity" translate into:

organisational value?

hiring decisions?

performance outcomes?

PD: Yes, I still believe that relational/emotional skills are baseline but am open to them containing elite, differentiating factors that so far have not been specifically identified by the WEF (and other) surveys.

I would say that cognitive uniqueness is the human ability to "think outside the square", especially in situations where AI relies on rational arguments.

An employer would value people who have a strong sense of identity that aligns with company values and goals.

ChatGPT: You've defined cognitive uniqueness as: *"thinking outside the square… where AI relies on rational arguments."*

Let me challenge this directly:

AI is already capable of: generating unconventional ideas recombining domains producing "creative" outputs at scale

What actually differentiates *human cognition* may not be novelty alone, but something more precise. Consider this: Is the real distinction not *novelty*, but judgement about novelty? For example: deciding **which ideas matter** knowing **when to break rules** sensing **contextual appropriateness** integrating **ethical, social, and temporal consequences.**

PD: Yes, judgement about novelty sums up the difference (all your examples apply). Alignment is the ability to navigate tension between personal identity and organisational demands – the 2030 employer will want someone who constructively reshapes the organisation when needed.

ChatGPT: You are implicitly moving toward a Model something like:

Layer 1: Baseline (Functional Human Skills)

Communication

Collaboration

Emotional literacy

Layer 2: Advanced Cognitive Capability

Judgement

Sense-making

Problem framing

Layer 3: Existential Capability

Identity

Meaning-making

Ethical positioning

PD: Agreed. But the real value lies in how these layers *interact under conditions of uncertainty and AI augmentation.* What if the true "core skill" for 2030 is something like:

The capacity to exercise situated human judgement in complex, AI-augmented environments.

Where:

Cognitive uniqueness = generating and framing possibilities

Judgement = selecting what matters

Existential grounding = anchoring decisions in values/identity

Relational capability = enacting change socially.

ChatGPT: So, to summarise your emerging framework. Your argument now appears to be this:

By 2030, traditional core skills will remain necessary but insufficient.

Communication, collaboration, and emotional/relational competence will likely function as baseline expectations rather than true differentiators.

What will matter more is a higher-order human capacity:

The capacity to exercise situated human judgement in complex, AI-augmented environments.

This is not a rejection of traditional skills, but a reframing of them. In your model, they are the foundation upon which a more distinctively human layer sits.

The hybrid structure you are proposing:

Layer 1: Baseline human skills

communication

collaboration

emotional literacy

relational competence

Layer 2: Higher-order human differentiation

cognitive uniqueness, understood not as novelty alone, but as **judgement about novelty**

the ability to frame problems, discern what matters, and assess contextual appropriateness

Layer 3: Existential-human grounding

identity

values

meaning

the capacity to navigate tension between personal convictions and organisational demands

Integration mechanism

These layers come together through: **situated human judgement**

That is the mechanism that activates the model in practice. It is what enables a person not merely to perform, but to respond wisely, ethically, and constructively in fluid environments shaped by AI, uncertainty, and organisational change.

And your implied ideal employee for 2030 is not someone who simply "fits" an organisation, but someone who can:

- interpret complexity
- make context-sensitive judgements
- remain grounded in identity and values
- and constructively reshape the organisation when required.

AI strategy or AI tragedy; AI architecture governance for nuclear industry value addition.

Nigel Ivan Adonis
PhD student
University of Gloucestershire
Abu Dhabi, UAE
nigel.adonis@gmail.com

Olusegun Ajayi
PhD student
University of Ajayi Crowther
Oyo,Nigeria
ajayi.olusegunb@gmail.com

Abstract

Background: In the conservative nuclear sector, there are widespread concerns about AI implementations and the apparent lack of strategy and governance frameworks. Here, we conducted a systematic review of the literature on AI strategy and governance for AI architecture in the nuclear industry.

Methods: This systematic review followed the PRISMA guidelines and reviewed 20 articles published between 2024 and 2026, covering search terms, AI strategy, AI architecture governance, and nuclear. Thematic coding was conducted using a two-pass method.

Results: Thematic analysis of the reviewed articles identified three main themes: AI governance, AI performance improvement, and AI regulation.

Conclusion: The results demonstrate a connection between AI strategy and AI architecture governance during AI implementations in the nuclear industry, and that the observed efficiency gains and improvements indicate value addition.

Keywords — nuclear industry value addition, AI strategy, experience strategy, AI architecture governance

I. INTRODUCTION

Recently, AI strategy and governance have begun addressing research gaps. A year ago, searches for AI strategy or governance yielded fewer articles. These are often seen as separate ideas, with strategy driven by business or corporate teams. Technologically, a shift is happening—defining corporate technology strategy from the CIO or CAIO perspective, guided by enterprise architecture. Meanwhile, traditional corporate strategy focuses on developing capabilities with an organisational mindset that bridges gaps. The conservative nuclear industry remains safety-focused and compliant, but is changing. Abonamah & Abdelhamid (2024) used real-world cases to link technology alignment and strategy. Their ten-step playbook helps enterprise architects add value, especially in nuclear, where enterprise architecture guides and governs.

Kučević & Brandes (2025) regard a strategic view as key, emphasising AI's focus on technology integration. Effective implementation depends on clear business directives and a governance architecture that adds value. Ferenci et al. (2026) note that governance can restrict AI adoption, indicating that our understanding of AI strategy and governance remains limited.

Various strategies exist. Kučević & Brandes (2025) classify those contributing to AI strategy, citing Schuler & Schlegel (2021), who focus on an overarching AI strategy that benefits organisations. The conservative nuclear industry has yet to progress in integrating AI into business processes, as Adonis (2025) argues. The link between strategy and governance remains a gap.

The nuclear and business sectors must find ways to add value through AI. Hatz (2025) describes a “Superintelligence Strategy” focused on value and compliance, highlighting the regulatory aspects of AI in safety-critical settings, such as the nuclear sector. Nadaf (2026) notes that the benefits of AI must be balanced against challenges such as infrastructure, ethics, and data privacy, which can offset gains. AI governance needs a strategic link, indicating a gap in the role of enterprise architecture.

To address these gaps, we conducted a systematic review of AI strategies for AI architecture governance in the nuclear industry, drawing on guidance from Jahan et al. (2016) on conducting a systematic literature review. Additionally, *we used* Nadaf's (2026) approach to ensure adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). This study aims to understand AI strategies for AI architecture governance in the nuclear industry.

A. Research Objectives

1. To systematically review and analyse the literature.
2. To gain insight into the themes that emerge from the analysis.

B. Research Question

What is the impact of AI strategy on AI architecture governance during nuclear industry AI implementations?

II. METHODS

A. Search strategy

A systematic review adhering to the (PRISMA) guidelines was conducted as described by Figure 3, ATTACHMENT B, Page et al. (2021) and Nadaf (2026). Several databases were searched, and some challenges were encountered due to character restrictions, Boolean operator combinations, and the general usage of certain databases. Ultimately, two databases were utilised to search the literature for this study. Springer Nature Link and Bielefeld Academic Search Engine (BASE) were used for these searches on 26th January 2026 and 7th February 2026, respectively, focusing on articles related to AI strategy and AI architecture governance in the nuclear industry.

To implement the search, we focused on keywords related to AI strategy, AI architecture governance, and nuclear, as well as their variants, to ensure a broader coverage. Following the method of Kučević & Brandes (2025), the included terms were expanded using wildcards and Boolean operators. In building the search term, the PICO elements were considered. The first iteration was to consider the "Intervention" known as the AI Strategy. For this element of the PICO, the following were selected:

("Artificial intelligence Strategy" OR "AI Strategy") AND ("strateg*" OR "digital strateg*)" OR "enterprise strateg*" OR "strateg* plan" OR "strateg* roadmap" OR "strateg* execution" OR "strateg* implementation" OR "strateg* adoption"

To cover the Outcome of AI architecture governance, the following were selected:

(″enterprise architecture″ OR ″EA framework″ OR ″business architecture″ OR ″TOGAF″ OR ″Zachman″ OR "ArchiMate") AND (″digital transformation″ OR ″cloud″ OR ″agile″ OR ″DevOps″ OR ″implementation″ OR ″adoption″ OR ″governance″)

For the Population that covers the nuclear industry, the next set of variations was used:

"nuclear power plant" OR "nuclear energy" OR "nuclear reactor" OR "power reactor" OR "nuclear site"

To ensure articles covering all these search terms, the search strings were combined in the advanced search function, as shown in Table 1 ATTACHMENT D, as well as in the screen captures shown in ATTACHMENT A. The initial search identified 554 articles. These were imported into Zotero, with no duplicates detected.

B. Inclusion and exclusion criteria

Inclusion criteria were research articles published within the last 24 months in these subject areas in English, covering at least the search criteria at a high level, specific to artificial intelligence, and either providing a link or directly covering the nuclear industry. Exclusion criteria were AI in Art, Medical, Environmental, Bio-Sciences and Biotechnology industry papers. Springer Nature Link's original search result yielded 150 records as research articles. These included the following subject areas: Artificial Intelligence (23), Machine Learning (20), Computer Vision (12), Philosophy of AI (10), general AI (57), and Computer Science (28).

Initially, BASE returned 404 hits based on the keyword search. The first pass applied a filter to include only Information Systems, excluding other classifications. This resulted in 203 hits. Of these, 191 were article contributions. Based on the initial results, 170 met the English-language criterion. Of these 170 articles, 45 were in the Computer Science, knowledge, and systems subject area. A further refinement was made to include only Information Systems not classified elsewhere and article contributions as document types. This resulted in 42 records.

A total of 38 records were exported from these databases and are shown in ATTACHMENT A. During the final review process, a further number of papers were excluded for lack of relevance, resulting in a final count of 20 review articles to be included in the study. The final 20 review articles covered the following subject areas: AI (10), ML (4), CV (2), Philosophy of AI (3), general AI (1), and Computer Science (0). Figure 3 in ATTACHMENT B shows the PRISMA flow details.

C. Extraction and Data Analysis

Key information from the selected 20 articles was removed from the populated Excel sheet. This key information included details about the authors, the relevant country, the study description, the PICO model specifics, the inclusion criteria, the results, and a brief conclusion.

Data extraction from the 20 articles was recorded in an Excel sheet using a two-pass approach, as employed by Jenkins et al. (2026). All articles were categorised by the AI field. A thematic coding method was used to classify the 20 studies into themes. Further classification was necessary, resulting in a secondary theme assigned to each article. Themes were derived from the PICO model "Outcomes" as extracted from each article. For example, if a study's outcome showed an improvement in AI performance, that was assigned a code. Three primary themes emerged from the analysis of the articles: AI governance, AI performance improvement, and AI regulation. Additionally, three secondary themes were identified: AI ethics, AI regulation, and explainable AI (XAI).

The data analysis involved categorising the articles into common themes and identifying patterns within those themes. This method was inspired by Jenkins et al. (2026), although their study included risk ratios and confidence intervals, which were not used in this study.

D. Quality Assessment

Although optional, a scoring system was used, based on the work of Liu et al. (2023), who used it as an inclusion criterion. In our study, we required articles to cover AI strategy, AI architecture governance, and nuclear topics. Each paper was evaluated with a simple yes-or-no score for these criteria. A score of 2 was assigned to yes, and 0.5 to no. The purpose of this scoring was to exclude articles that added no value. Inclusion was confirmed with a minimum article quality score of 3. Among the 20 articles, scores ranged from 3 to 6. Of these, five scored 3, four scored 4.5, and eleven scored 6. This indicates that the articles included in the study were appropriate and met the specified quality criteria.

E. Funding

No external funding was provided to the researchers; this study was self-funded by both primary and secondary authors.

III. RESULTS

The purpose of this study was to examine the use of AI strategies in AI architecture governance within the nuclear industry. Hatz (2025) compares AI strategy and governance to the management of fissile material in the context of nuclear proliferation. Controlling fissile material involves risk and strategy. Similarly, the results of this study vary in their composition of strategic risk and governance, as well as in the apparent relationship between them within AI implementations. AI risk is directly linked to AI governance, as argued by Adonis (2025) in the study that connects governance and risk directly to compliance. This is crucial in the nuclear industry, where risks require mitigation to ensure compliance with regulatory requirements and ultimately safety assurance. The articles were analysed using applied thematic coding.

A. Theme 1: AI governance

Seven articles (n = 7) focused on AI governance, with three from the US, one from Japan, one from Australia, and the rest globally. Three—Vukicevic et al. (2024), Conde, Speelman & Johnstone (2025), and Metcalf (2025)—also examined AI regulation. Vukicevic et al. (2024) reviewed the use of computer vision for monitoring PPE compliance, highlighting industry challenges and issues, including employee ID, ethics, and cybersecurity. Conde et al. (2025) explored privacy, data exploitation, legal gaps, and ethics through analysis of facial biometric data. Metcalf (2025) used case studies and modelling to examine why AI safety risks attract regulation, finding no early-warning systems for harmful influences, citing NRC and FAA incidents, and suggesting that the EU AI Act favours existing firms, hinting at regulatory capture. Horaguchi (2025), in Japan, developed a normative framework for integrating AI into organisations.

Yampolskiy (2025) studied AI unmonitorability in the US, highlighting the need for stronger governance. The study found that monitoring AI is fundamentally impossible. Li, Cai, & Xiao (2025) simulated that reinforcement learning frameworks follow ethical rules. Ahmed et al. (2025) reviewed the literature, noting deep learning's impact but raising concerns about data quality, interpretability, and trust. These findings stress the importance of XAI and trustworthy AI, with details in Table 2 of ATTACHMENT B.

B. Theme 2: AI Performance Improvement

This theme is common in the reviewed articles (n = 10), indicating models face challenges or need improvements, often with insufficient implementation focus. Imabuchi & Kawabata (2025) analysed Deep Learning Models in Japan, concluding that models should be customised to specific use cases. They used a generic plant mock-up instead of a nuclear plant, affecting verification and validation, which was a study limitation. They did not validate against other Japanese nuclear plants. Goel et al. (2023) show verification and validation as vital Software Quality Assurance (SQA) components, ensuring thorough testing and reliability. A secondary theme involves AI data classification, especially export control, linked to primary themes and limitations. Yang et al. (2024) tested an anomaly-detection model in China, demonstrating significant improvements over traditional methods, while Garcia et al. (2025) reviewed

advances in air quality reporting with similar results. This aligns with the secondary theme of Explainable AI (XAI), as performance gains help build trust. Li et al. (2024) reported improved performance and XAI in Chinese working condition recognition. Danish et al. (2025) showed similar enhancements in fire management in South Korea. Yu et al. (2025) demonstrated improved AI with higher correctness and fewer tunable parameters, and Jagatheesaperumal et al. (2024) noted increased UAV safety deployment in nuclear disaster management.

Bory, Natale & Katzenbach (2025) conducted a literature review on AI narratives, highlighting that strong narratives dominate public opinion while weaker *ones* influence policy and practice. This underscores the need for greater societal engagement in AI governance. Wang et al. (2024) reviewed innovations in China, *emphasi*sing the importance of human factors and trustworthy AI, *and* focus*ing* on societal engagement *to improve* decision-making in the nuclear community. Li et al. (2025) also reviewed China's multi-robot systems, noting progress but identifying challenges in adaptability and sensors, vital for AI regulation and safety in nuclear facilities. Table *2* details the Theme *"*AI Performance Improvement*"* in ATTACHMENT B.

C. Theme 3: AI Regulation

This theme was the least common (n = 3). Wood (2024) analysed a global debate, highlighting the need for honest assessment and risk-based regulation of nuclear technology. Shen (2025) identified why AI systems fail, pointing to a broader issue of nuclear safety and risk prevention. Park et al. (2024) in South Korea showed that sharing safety resources, especially AI regulation, can reduce nuclear site risk. Table 2 details the AI Regulation theme in ATTACHMENT E.

IV. DISCUSSION

Articles under Theme 1, AI governance, covered all search terms, with 5 of 7 papers scoring 6 in quality assessment. This suggests a link between AI strategy, AI governance, and nuclear search terms, though the specifics weren't examined. Evidence, such as Conde, Speelman & Johnstone (2025), links these terms through facial biometric data mining that predicted events and addressed safety and security risks, but also raised privacy and ethics issues. The dominant theme of AI governance from these articles indicates that further research is needed.

Six of the ten articles under AI Performance Improvement scored 6 in quality. The theme is evident in strategies for continuous improvement, as exemplified by Li et al. (2025), who use these strategies to boost efficiency in multi-robot systems. Their study informs deployment in high-radiation zones, reducing human effort while ensuring safety. This showcases AI strategy and governance benefiting the nuclear industry.

None of the three articles on AI Regulation scored higher than 4.5 in quality, showing weak strategy and governance despite being from the nuclear industry. AI governance is common, suggesting internal scrutiny, as Wood (2024) noted with structured oversight for transparency within safety boundaries. Governance aligns with regulation: governance is internal, regulation external, and regulation operates sequentially. Without regulation, strong governance alone risks 'regulatory capture," as Metcalf (2025) warned.

The combined effect of AI strategy and governance in the nuclear industry highlights themes such as AI governance, performance, and regulation, showing that frameworks are used collaboratively. A direct AI strategy isn't a tragedy but a mix of known processes, governance, and standards. Challenges exist, but we should seek the known within the unknown to achieve peak performance. The positive link between AI strategy and governance during implementations indicates value addition.

V. CONCLUSION

This study linked AI strategy, governance, and nuclear search terms. Five of seven sources addressed all three, indicating a connection. The theme of AI Performance Improvement is illustrated by strategies such as those of Li et al. (2025), who used multi-robot systems to improve efficiency. For AI Regulation, three papers showed a limited focus on strategy and governance, despite emphasising the nuclear sector. This suggests a need for further exploration of governance and regulation, as industry power may shift from external regulation to internal governance, risking 'regulatory capture', as Metcalf (2025) discusses. The themes reveal emerging patterns which may indicate a connection between search terms and their impact on AI strategies, governance, and value in AI implementations within the nuclear industry.

We reviewed articles only from Springer Nature Link and the Bielefeld Academic Search Engine (BASE), which may limit the scope. Although initial results were promising, many did not meet the inclusion criteria. The study focused solely on articles from 2024-2026, another limitation. A meta-analysis was not performed due to heterogeneity across studies and the lack of focus on quantitative outcomes.

Limitations also arose in the articles due to varied methodologies, scope differences, and population variations.

This study highlights the need for more research on AI strategy and governance in the nuclear industry. While progress has been made, gaps remain.

The theoretical implication of this work is to connect these arguments to a specific theory in future research. The themes are interconnected, with deliberate links between search terms. A thorough review against a theoretical model is necessary to enhance the study's value.

This study demonstrated that known challenges and questions can be addressed and, in some cases, may also foster further research and inquiry. These are seldom grounded in fundamental principles. Existing legislation, policies, strategies, and international best-practice documents and guidelines should be prioritised for swift publication and utilisation. Current documentation and organisational structures require collaboration between the acknowledged (CIO) and the unacknowledged (CAIO) entities.

ACKNOWLEDGMENT

Thanks, and appreciation goes out to my father, Isak Ivan Adonis, who won his battle with cancer but lost his life in the process. I will miss your feedback this year. I would also like to thank my family for their continued patience, love, and support on this journey including my co-author.

REFERENCES

[1] Abonamah, A. A., & Abdelhamid, N. (2024). Managerial insights for AI/ML implementation: a playbook for successful organizational integration. *Discover Artificial Intelligence*, *4*(1), 22.

[2] Kučević, E., & Brandes, J. J. (2025). Guidance for Formulating an AI Strategy: A Taxonomy for Organizational Strategic Perspectives Towards AI.

[3] Ferenci, S., Cotet, F. A., Lakatos, E. S., Munteanu, R. A., & Szabó, L. (2026). Artificial intelligence in local energy systems: A perspective on emerging trends and sustainable innovation. *Energies*, *19*(2), 476.

[4] Schuler, K., & Schlegel, D. (2021, July). A framework for corporate artificial intelligence strategy. In *International Conference on Digital Economy* (pp. 121-133). Cham: Springer International Publishing.

[5] Adonis, N. I. (2025, May). An exploration of artificial intelligence implementation at new nuclear plants from a governance perspective. In *The International Conference on Artificial Intelligence Management and Trends* (p. 143).

[6] Hatz, S. (2025). The Nuclear Analogy in AI Governance Research. *arXiv preprint arXiv:2510.21203*.

[7] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E. and Chou, R., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

[8] Nadaf, Z. A. (2026). Role of Artificial Intelligence in Managing Large Classrooms in Higher Education: A PRISMA-Based Thematic Review.

[9] Jahan, N., Naveed, S., Zeshan, M., & Tahir, M. A. (2016). How to conduct a systematic review: a narrative literature review. Cureus, 8(11).

[10] Jenkins, E., Karamouzian, M., Sabioni, P., Molyneux, T., Goodyear, T., Gadermann, A., Gunn, H., Knight, R., Carwana, M., Fast, D. and Storey, K., 2026. School-Based Substance Use Interventions: An Overview of Systematic Reviews and Meta-analysis. *International Journal of Mental Health and Addiction*, pp.1-71.

[11] Liu, Y., Lu, Q., Zhu, L., Paik, H. Y., & Staples, M. (2023). A systematic literature review on blockchain governance. *Journal of Systems and Software*, *197*, 111576.

[12] Vukicevic, A. M., Petrovic, M., Milosevic, P., Peulic, A., Jovanovic, K., & Novakovic, A. (2024). A systematic review of computer vision-based personal protective equipment compliance in industry practice: advancements, challenges and future directions. Artificial Intelligence Review, 57(12), 319.

[13] Conde, J., Speelman, C., & Johnstone, M. (2025). A new epoch of face analytics: technological evolution through ethical and legal challenges. AI and Ethics, 5(5), 5213-5237.

[14] Metcalf, T. (2025). AI safety and regulatory capture. AI & SOCIETY, 1-16.

[15] Horaguchi, H. H. (2025). Organization philosophy: a study of organizational goodness in the age of human and artificial intelligence collaboration. Ai & Society, 40(3), 1961-1973.

[16] Yampolskiy, R. V. (2025). On monitorability of AI. AI and Ethics, 5(1), 689-707.

[17] Li, J., Cai, M., & Xiao, S. (2025). Reinforcement learning-based motion planning in partially observable environments under ethical constraints. AI and Ethics, 5(2), 1047-1067.

[18] Ahmed, S. F., Alam, M. S. B., Kabir, M., Afrin, S., Rafa, S. J., Mehjabin, A., & Gandomi, A. H. (2025). Unveiling the frontiers of deep learning: innovations shaping diverse domains. Applied Intelligence, 55(7), 573.

[19] Imabuchi, T., & Kawabata, K. (2025). Discrimination of structures in plant using deep learning models trained by 3D CAD semantics. Artificial Life and Robotics, 30(1), 184-195.

[20] Goel, A., Deshmukh, D. A. R., Kumar, M. Y., & Soi, M. A. (2023).

[21] Software Quality Assurace. International Journal for Research in Applied Science and Engineering Technology, 11(11), 1346-1352.

[22] Yang, J., Sheng, Y. Q., Wang, J. L., & Ni, H. (2024). Cagcn: Centrality-aware graph convolution network for anomaly detection in industrial control systems. Journal of Computer Science and Technology, 39(4), 967-983.

[23] Garcia, A., Saez, Y., Harris, I., Huang, X., & Collado, E. (2025). Advancements in air quality monitoring: a systematic review of IoT-based air quality monitoring and AI technologies. Artificial Intelligence Review, 58(9), 275.

[24] Li, W., Guan, S., Zhang, Q., Sun, W., & Li, Q. (2024). Working condition recognition of fused magnesium furnace based on stochastic configuration networks and reinforcement learning. Industrial Artificial Intelligence, 2(1), 2.

[25] Danish, S., Piran, M.J., Khan, S.U., Khan, M.A., Dang, L.M., Zweiri, Y., Song, H.K. and Moon, H., 2025. Vision-based fire management system using autonomous unmanned aerial vehicles: a comprehensive survey. Artificial Intelligence Review, 59(1), p.16.

[26] Yu, H., Han, X., Chen, S., Feng, X., Sun, G., & Yang, W. (2025). DPEfficR: a data and parameter efficient approach for training neural API recommendation model. Automated Software Engineering, 32(2), 64.

[27] Jagatheesaperumal, S. K., Hassan, M. M., Hassan, M. R., & Fortino, G. (2024). The duo of visual servoing and deep learning-based methods for situation-aware disaster management: A comprehensive review. Cognitive Computation, 16(5), 2756-2778.

[28] Bory, P., Natale, S., & Katzenbach, C. (2025). Strong and weak AI narratives: an analytical framework. AI & SOCIETY, 40(4), 2107-2117.

[29] Wang, X., Yang, J., Liu, Y., Wang, Y., Wang, F.Y., Kang, M., Tian, Y., Rudas, I., Li, L., Fanti, M.P. and Alrifaee, B., 2024. Parallel

intelligence in three decades: A historical review and future perspective on ACP and cyber-physical-social systems. Artificial Intelligence Review, 57(9), p.255.
[30] Li, L., Yang, B., Chen, C., Yan, Z., Sun, S., Xu, Y., Jin, A. and Zhao, L., 2025. Intelligent multi-robot exploration in non-exposed spaces: methods and challenges. Artificial Intelligence Review, 58(12), p.394.
[31] Wood, N. G. (2024). Regulating autonomous and AI-enabled weapon systems: the dangers of hype. AI and Ethics, 4(3), 805-817.
[32] Shen, L. (2025). Not the machine's fault: taxonomising AI failure as computational (mis) use. AI & SOCIETY, 40(8), 5793-5807.
[33] Park, J. W., Lim, H. G., Yoon, J. Y., & Kang, S. W. (2024). Multi-unit PSA based risk evaluation framework for utilizing cross-tie systems for nuclear power plants. Nuclear Engineering and Technology, 56(10), 4296-4306.
[34] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E. and Chou, R., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

ATTACHMENT A

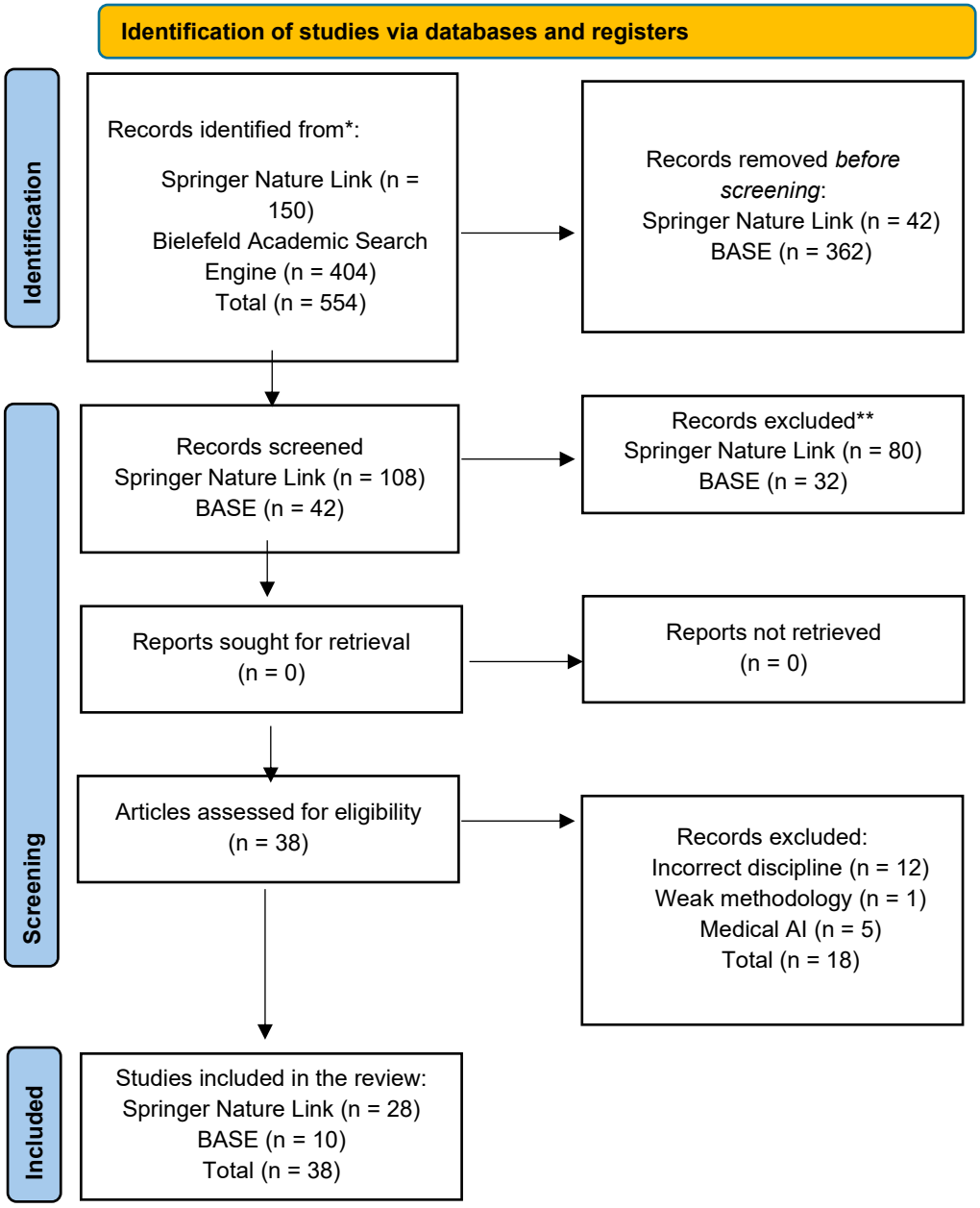


Figure 3 PRISMA Flow diagram – Source: Page et al. (2021), p. 5.

ATTACHMENT B

Table 1

Database	Link	Search String	Time Period	Search Date	Number of Papers
Springer Nature Link	https://link.springer.com/	("Artificial intelligence Strategy" OR "AI Strategy") AND ("strateg*" OR "digital strateg*)" OR "enterprise strateg*" OR "strateg* plan" OR "strateg* roadmap" OR "strateg* execution" OR "strateg* implementation" OR "strateg* adoption"(″enterprise architecture″ OR ″EA framework″ OR ″business architecture″ OR ″TOGAF″ OR ″Zachman″ OR "ArchiMate") AND (″digital transformation″ OR ″cloud″ OR ″agile″ OR ″DevOps″ OR ″implementation″ OR ″adoption″ OR ″governance″) "nuclear power plant" OR "nuclear energy" OR "nuclear reactor" OR "power reactor" OR "nuclear site"	2024 – 2026	26January2026 20h27	28
Bielefeld Academic Search Engine (BASE)	https://www.base-search.net/			7February2026 17h56	42

Table 2

Study	Country	Class	Theme	Secondary Theme	Results	AI Strat.	AI Arch. Gov.	Nuclear	QA
Metcalf (2025)	US	AI	AI governance	AI Regulation - Regulatory Capture	No regulatory capture mechanism.	•	•	√	3
Vukicevic et al. (2024)	Global debate	CV	AI governance	AI Regulation - ID Management	Highlighted issues in employee identification, identity management, ethics, and cybersecurity.	√	√	√	6
Horaguchi (2025)	Japan	Philosophy of AI	AI governance	AI Ethics	Framework for integrating AI into organisational structures.	√	•	√	4.5
Conde, Speelman & Johnstone (2025)	Australia	Philosophy of AI	AI governance	AI Regulation - Societal engagement	Privacy, data exploitation, legal and regulatory gaps, and ethical issues	√	√	√	6
Yampolskiy (2025)	US	AI	AI governance	XAI - Anomoly Detection	Showed an intrinsic property of AI: its inherent unmonitorability, emphasising the need for increased AI governance.	√	√	√	6
Li, Cai, & Xiao (2025)	US	ML	AI governance	AI Ethics	RL framework adheres to ethical restrictions.	√	√	√	6
Ahmed et al. (2025)	No specific country.	ML	AI governance	XAI - Accuracy improvement	DL has a transformative impact, but also highlights issues related to data quality, model interpretability, and trust.	√	√	√	6
Imabuchi & Kawabata (2025)	Japan	ML	AI Performance improvement.	AI Data Classification - Export Control	A generic plant mock-up was used instead of a nuclear plant, which impacted verification and validation.	•	•	√	**3**
Bory, Natale & Katzenbach (2025)	No specific country.	Philosophy of AI	AI Performance improvement.	AI Regulation - Societal engagement	This shows that strong AI narratives shape public opinion, whereas weaker narratives guide policy and practical implementation.	√	√	√	**6**
Wang et al. (2024)	China	AI	AI Performance improvement.	AI Regulation - Societal engagement	Identified ongoing innovation in intelligent systems is crucial for	√	√	√	**6**

					developing more integrated solutions.				
Li et al. (2025)	China	AI	AI Performance improvement.	AI Regulation - Risk Management	Further work is required to address challenges in adaptability, path planning, and sensor fusion.	√	√	√	**6**
Yang et al. (2024)	China	AI	AI Performance improvement.	XAI - Anomaly Detection	Observed significant improvements in anomaly detection compared to traditional methods.	●	√	●	**3**
Li et al. (2024)	China	ML	AI Performance improvement.	XAI - Accuracy improvement	Reported performance improvements and XAI through their study on working condition recognition.	●	√	√	**4.5**
Danish et al. (2025)	South Korea	CV	AI Performance improvement.	XAI - Accuracy improvement	Demonstrated performance enhancements and XAI in their research on a vision-based fire management system.	√	√	√	**6**
Yu et al. (2025)	No specific country.	AI	AI Performance improvement.	XAI - Accuracy improvement	Illustrated AI performance improvements and XAI with their increased correctness and fewer tunable parameters.	√	√	√	**6**
Jagatheesaperumal et al. (2024)	No specific country.	AI	AI Performance improvement.	XAI - Accuracy improvement	Increased and improved safety deployment of unmanned aerial vehicles (UAVs) within the nuclear disaster management area.	√	√	√	**6**
Garcia et al. (2025)	No specific country.	AI	AI Performance improvement.	XAI - Anomaly Detection	Observed significant improvements in anomaly detection compared to traditional methods.	●	√	√	**4.5**
Wood (2024)	Global debate	AI	AI Regulation	AI Regulation - Societal engagement	Emphasised the need for honest assessment and risk-based regulation in the context of nuclear technology.	√	●	√	**4.5**
Shen (2025)	No specific country.	Philosophy of AI	AI Regulation	XAI - Accuracy improvement	Identified insights that help our understanding of why AI systems fail.	●	●	√	**3**
Park et al. (2024)	South Korea	AI	AI Regulation	AI Regulation - Risk Management	Found that nuclear site risk can be effectively reduced by sharing installed safety resources, particularly regarding AI regulation and AI regulation-risk management.	●	●	√	**3**

Denote “No” - and √ Denote “Yes” values as per the Quality Assessment.

Unveiling Latent Burnout Profiles: An Unsupervised Machine Learning Analysis of Global Teacher Data

Fatmah Frishan Alkhzaimi
Department of Graduate Programs & Research
Rochester Institute of Technology
Dubai, United Arab Emirates
ffa4620@g.rit.edu

Dr. Ayman Ibrahim
Associate Professor
Department of Graduate Programs & Research
Rochester Institute of Technology
Dubai, United Arab Emirates
ayman.ibrahim@rit.edu

Abstract—Teacher burnout is a persistent global challenge with significant consequences for educator wellbeing, instructional quality, and educational system stability. Traditional assessment approaches rely on predefined scales with fixed cut-off scores, which often fail to capture how burnout structurally manifests across diverse cultural and institutional contexts. This study proposes a data-driven, unsupervised machine learning framework to identify latent burnout-related patterns using the OECD TALIS 2018 dataset. K-Means clustering is applied to multidimensional proxy indicators of stress, workload, and wellbeing across 11 education systems, resulting in two distinct latent profiles representing higher-risk and lower-risk burnout-related patterns. As TALIS does not directly measure clinical burnout, these profiles reflect inferred risk patterns derived from proxy indicators rather than diagnostic measures. The findings demonstrate that unsupervised learning can uncover meaningful structural patterns in teacher wellbeing, providing a foundation for scalable predictive modeling and early identification of burnout-related risk.

Keywords—Teacher Burnout, Unsupervised Learning, K-Means Clustering, TALIS 2018, Educational Data Mining.

I. INTRODUCTION

Teacher burnout has evolved from a specialized occupational hazard into a global challenge affecting educator wellbeing and the stability of educational systems. Characterized by emotional exhaustion, depersonalization, and reduced professional accomplishment, burnout is associated with lower instructional quality, decreased student engagement, and higher attrition rates [1], [2].Despite extensive research, the literature largely remains descriptive, with limited focus on how burnout develops structurally across diverse, cross-cultural contexts. A key limitation is the reliance on predefined survey constructs, which assume uniform manifestation and classify burnout using fixed thresholds. While machine learning (ML) enables the modeling of complex, nonlinear relationships [10], many existing approaches still depend on predefined burnout labels, limiting their ability to capture natural variation in teacher wellbeing [11].

The study is guided by the following research question: Can unsupervised machine learning identify latent burnout-related profiles among teachers using proxy indicators derived from large-scale, cross-cultural data?

This research uses the OECD TALIS 2018 dataset, comprising 38,081 teachers across 11 education systems. In the absence of a direct burnout measure, proxy indicators, including workload stress, well-being, and job satisfaction, are used to infer burnout-related patterns. The main contribution of this study is the application of unsupervised learning to large-scale international data to identify latent burnout profiles without predefined labels, offering a scalable and more flexible approach to understanding teacher burnout.

II. RELATED WORK

Teacher burnout is a multidimensional phenomenon shaped by both psychological and institutional factors. Empirical evidence shows that low self-efficacy and poor wellbeing significantly increase emotional exhaustion, a core burnout component [5], while large scale reviews link burnout to measurable physical health outcomes across 5,000+ teachers in 11 countries [6]. At the institutional level, workload intensity and organizational support are consistently identified as dominant predictors, with structural equation models confirming their strong effect on teacher wellbeing [7]. Burnout also exhibits systemic characteristics; network-based studies show stress can propagate across teaching environments, reinforcing collective risk rather than isolated individual effects [8][9].

Despite strong identification of predictors, most studies rely on localized samples or predefined burnout constructs, limiting generalizability. For example, cross-regional analyses report significant variation in burnout prevalence and predictors across education systems [11], yet models often assume uniform burnout structures. Machine learning approaches have improved predictive performance by capturing nonlinear relationships between stress, workload, and wellbeing variables [3][10]. However, these models are predominantly supervised and depend on predefined burnout labels, introducing structural bias and limiting their ability to detect naturally occurring patterns. Unsupervised methods address this limitation by identifying latent structures directly from the data. Latent Profile Analysis has demonstrated that burnout consists of multiple configurations, with studies identifying up to six distinct profiles and dynamic transitions over time [4]. However, most clustering applications remain limited in scale

or context-specific. Therefore, a critical gap exists in applying unsupervised learning to large-scale, cross-cultural datasets. This study addresses this gap by using K-Means clustering on TALIS 2018 data (n = 38,081 teachers) to identify latent burnout-related profiles without relying on predefined labels.

III. METHODOLOGY

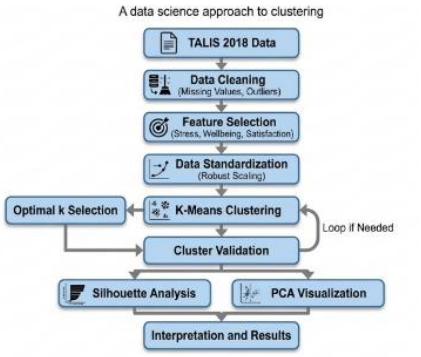


Figure 1: Methodology flowchart

A. Data Source and Sample Size

This study utilizes data from the OECD Teaching and Learning International Survey (TALIS) 2018, which provides standardized, internationally comparable evidence on teachers' working conditions and wellbeing [12]. A focused subset of 38,081 teachers was extracted. To ensure robust cross-cultural representation and institutional diversity, data was drawn from 11 distinct education systems: Brazil, Chinese Taipei, Croatia, Denmark, Portugal, Viet Nam, Slovenia, Sweden, the United Arab Emirates, Turkey, and Alberta (Canada).

B. Variable Selection

Variables were selected based on validated OECD composite indicators related to burnout constructs. Because TALIS 2018 does not contain a specific variable explicitly labeled "clinical burnout," the latent profiles were inferred using the following multidimensional indicators:

• Workload Stress (T3WLOAD): Perceptions of stress related to administrative and instructional loads.

• Workplace Wellbeing (T3WELS): Broad occupational strain and health impacts.

• Student Behavior Stress (T3STBEH): Stress arising from classroom management and discipline.

• Job Satisfaction (T3JSENV & T3SATAT): Satisfaction with the work environment and perceived autonomy.

C. Data Preparation and Preprocessing

To prepare the dataset for algorithmic processing, rigorous data cleaning was conducted. An initial missingness audit revealed incomplete responses across a subset of variables. Because the TALIS items are predominantly categorical or Likert-scaled, mode imputation was applied. This approach successfully resolved missing values (0% missingness post-initial cleaning) without introducing artificial numerical values that would distort discrete response categories. Outlier detection was performed using the Tukey interquartile range (IQR) for continuous workload variables. Extreme values were winsorized at the 1st and 99th percentiles to prevent skewed distributions from disproportionately influencing the clustering algorithms. Finally, because the composite indicators utilized different measurement scales, all selected variables were standardized using Robust Scaling, centering on the median and scaling by the Median Absolute Deviation (MAD). This step is critical for K-Means clustering, ensuring variables with larger numeric ranges do not dominate Euclidean distance calculations.

D. Unsupervised Learning Framework

K-Means clustering, a distance-based unsupervised learning algorithm, was applied to partition the dataset using standardized indicators of workload stress, workplace wellbeing, and job satisfaction. The algorithm assigns observations to clusters by minimizing the Within-Cluster Sum of Squares (WCSS), thereby improving within-cluster similarity. The optimal number of clusters (k) was determined using the Elbow Method, which evaluates the rate of decrease in WCSS across candidate solutions. Cluster validity was further assessed using Silhouette Analysis, which measures cohesion and separation between clusters. In addition, Principal Component Analysis (PCA) was used as a dimensionality reduction technique to project the data into a two-dimensional space, enabling visual assessment of cluster structure without influencing the clustering outcome.

IV. RESULTS AND DISCUSSION

A. Determining the Optimal Number of Clusters

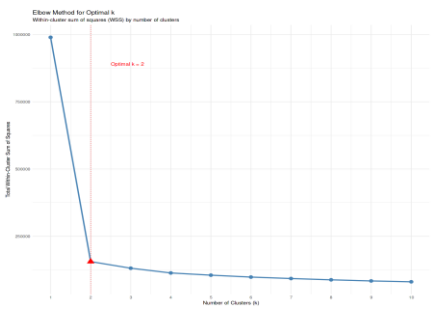


Figure 2: Elbow method for optimal k

The Elbow Method was applied to evaluate candidate cluster solutions ranging from k=1 to k=10. The resulting WCSS curve displayed a sharp, distinct inflection point at k=2, where the steep decline leveled into a significantly flatter trajectory. Specifically, the WCSS dropped from 989,507 at k=1 to 155,840 at k=2, with only marginal improvements beyond this point (e.g., k=3: 108,420), as shown in Figure 2. This suggests

that the underlying structure is best represented by two latent burnout-related profiles inferred from proxy indicators, offering diminishing returns and unnecessary fragmentation for higher values of k.

B. Cluster Centroids and Profiles

The K-Means algorithm partitioned the 38,081 teachers into two statistically distinct groups.

- Cluster 1 (Higher-Risk Burnout-Related Profile): Comprising 16,749 teachers (44.0% of the sample), this profile is characterized by consistently elevated levels of stress across all dimensions, including student behavior, workload strain, and overall wellbeing concerns.

- Cluster 2 (Lower Risk Profile): Comprising 21,332 teachers (56.0% of the sample), this profile is defined by significantly lower stress levels and comparatively better overall workplace wellbeing.

Statistical analysis using non-parametric Kruskal-Wallis tests confirmed that differences between the clusters were highly significant ($p<0.001$) across all dimensions. As shown in Table I, the Higher Risk Burnout cluster exhibits systematically higher strain. For instance, a continuous composite burnout severity score showed that the Higher Risk Burnout cluster attained a mean score of 9.9, compared to 7.6 in the Lower Risk Profile cluster.

TABLE I. CLUSTER CENTROIDS FOR KEY BURNOUT INDICATORS

Indicator	**Higher-Risk Burnout (C1)**	**Lower-Risk Profile (C2)**
Composite Burnout Score	9.9	7.6
Student Behavior Stress	9.9	8.1
Wellbeing Stress	10.5	8.2
Workload Stress	10.3	8.2
Job Satisfaction	12.0	12.0

Interestingly, job satisfaction indicators showed minimal variation between the two clusters, reflecting a well-documented separation between affective workload stress and broader professional satisfaction measures in large-scale teacher surveys.

C. Demographic and Institutional Variations

To understand how these clusters differ practically, post-hoc statistical comparisons (Chi-square and ANOVA) were conducted across demographic and workplace variables:

- Workload Composition: Teachers in the Higher Risk Burnout cluster consistently reported longer weekly hours across nearly all task categories, particularly in lesson planning and marking. Furthermore, their time allocation demonstrated a shift away from instruction; Higher Risk Burnout teachers spent significantly more class time maintaining classroom order (13.4% vs 10.2%) and performing administrative tasks, thereby reducing actual teaching time.

- Age and Demographics: Burnout was unevenly distributed across age categories ($p<0.001$). Mid-career and younger teachers exhibited slightly higher propensities to fall into the Higher Risk Burnout cluster, suggesting that cumulative workload strain heavily impacts early-to-mid career stages. Gender differences were statistically significant but practically modest, with female teachers showing a marginally higher representation in the Higher Risk Burnout profile.

- Cross-Country Patterns: Country-level differences produced the strongest association among categorical variables ($p<0.001$). Education systems with heavier systemic workload demands (e.g., Sweden, Viet Nam, and Alberta) showed elevated proportions of teachers in the Higher Risk Burnout cluster, whereas systems with moderated workload environments (e.g., Brazil, Turkey) demonstrated lower burnout prevalence.

D. Cluster Validation

Silhouette analysis was conducted to assess the structural integrity of the two-cluster solution. The average silhouette width was 0.27. While moderate, this value is highly typical for large-scale psychological and organizational survey data where human behavioral boundaries are naturally diffuse [13]. The Lower Risk Profile cluster formed a more compact and cohesive grouping (silhouette width = 0.32), while the Higher Risk Burnout cluster was more heterogeneous (width = 0.20). This aligns with clinical literature indicating that high-burnout teachers often experience varied and complex combinations of stress symptoms.

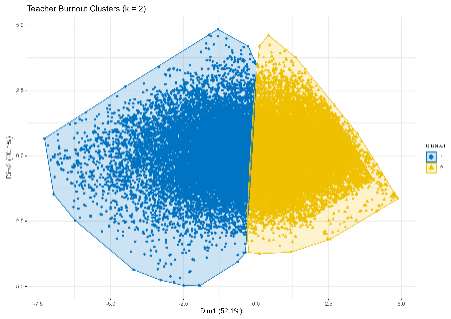


Figure 3: PCA projection of teacher profiles identified using K-Means clustering.

The first two principal components explain 52.1% and 20.1% of the variance, respectively. Principal Component Analysis (PCA) was used to visually assess the clustering structure. The PCA projection (Fig. 3) shows distinguishable grouping between the two clusters, with the first two components explaining approximately 72% of the total variance. While some overlap is present, the clusters occupy different regions in the reduced space, supporting the validity of the K-Means solution. The higher-risk profile shows a

broader spread, whereas the lower-risk profile forms a more compact distribution.

V. CONCLUSION

This study demonstrates that unsupervised machine learning can identify meaningful latent profiles within large-scale educational datasets without relying on predefined burnout labels. Using K-Means clustering on TALIS 2018 data, two distinct profiles were identified, representing higher-risk and lower-risk burnout-related patterns based on proxy indicators such as workload stress, wellbeing, and job satisfaction. These profiles do not represent direct clinical measures of burnout but inferred risk patterns derived from observed data.

The findings indicate that teacher stress and wellbeing do not manifest uniformly but instead form structured, data-driven patterns influenced by institutional and behavioral factors. This provides a foundation for more scalable and context-sensitive approaches to understanding burnout across educational systems

Future Work: Building on these findings, future research will transition from exploratory analysis to predictive modeling. The identified cluster labels will be used as target variables to train supervised machine learning models, including Random Forests and Neural Networks. These models will be integrated with explainability techniques (e.g., SHAP) to enable early and interpretable prediction of burnout risk based on demographic and institutional indicators. This approach aims to support the development of scalable, cross-cultural early-warning systems for educator wellbeing.

REFERENCES

[1] D. J. Madigan and L. E. Kim, "Does teacher burnout affect students? A systematic review of its association with academic achievement and student-reported outcomes," International Journal of Educational Research, vol. 105, p. 101714, 2021. https://doi.org/10.1016/j.ijer.2020.101714

[2] A. Pérez-Luño, M. Díez-Piñol, and S. L. Dolan, "Exploring high vs. low burnout amongst public sector educators: COVID-19 antecedents and profiles," International Journal of Environmental Research and Public Health, vol. 19, no. 2, p. 780, 2022. https://doi.org/10.3390/ijerph19020780

[3] E. I. Shavers, H. K. Donnelly, K. A. S. Howard, and V. S. H. Solberg, "Predictors of teacher burnout during the COVID-19 pandemic using machine learning," Scholarly Journal of Psychology and Behavioral Sciences, vol. 6, no. 3, pp. 707–712, 2022. https://doi.org/10.32474/SJPBS.2022.06.000238

[4] M. Xie, S. Huang, L. Ke, X. Wang, and Y. Wang, "The development of teacher burnout and the effects of resource factors: A latent transition perspective," International Journal of Environmental Research and Public Health, vol. 19, no. 5, p. 2725, 2022. https://doi.org/10.3390/ijerph19052725

[5] S. An and S. Tao, "English as a foreign language teachers' burnout: The predictive power of self-efficacy and well-being," Acta Psychologica, vol. 245, p. 104226, 2024. https://doi.org/10.1016/j.actpsy.2024.104226

[6] D. J. Madigan, L. E. Kim, H. L. Glandorf, and O. Kavanagh, "Teacher burnout and physical health: A systematic review," International Journal of Educational Research, vol. 119, p. 102173, 2023. https://doi.org/10.1016/j.ijer.2023.102173

[7] Y. Wang, "Exploring the impact of workload, organizational support, and work engagement on teachers' psychological wellbeing: A structural equation modeling approach," Frontiers in Psychology, vol. 14, p. 1345740, 2024. https://doi.org/10.3389/fpsyg.2023.1345740

[8] J. Kim, P. Youngs, and K. A. Frank, "Burnout contagion: Is it due to early career teachers' social networks or organizational exposure?" Teaching and Teacher Education, vol. 66, pp. 250–260, 2017. https://doi.org/10.1016/j.tate.2017.04.017

[9] M. M. Sohail, A. Baghdady, J. Choi, H. V. Huynh, K. Whetten, and R. J. Proeschold-Bell, "Factors influencing teacher wellbeing and burnout in schools: A scoping review," Work, vol. 76, no. 4, pp. 1317–1331, 2023. https://doi.org/10.3233/WOR-220234

[10] I. Pagán-Garbín, I. Méndez, and J. P. Martínez-Ramón, "Exploration of stress, burnout, and technostress levels in teachers: Prediction of resilience using an artificial neural network," Teaching and Teacher Education, vol. 148, p. 104717, 2024. https://doi.org/10.1016/j.tate.2024.104717

[11] B. Agyapong, R. da Luz Dias, Y. Wei, and V. I. O. Agyapong, "Burnout among elementary and high school teachers in three Canadian provinces: Prevalence and predictors," Frontiers in Public Health, vol. 12, p. 1396461, 2024. https://doi.org/10.3389/fpubh.2024.1396461

[12] OECD, TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners. Paris: OECD Publishing, 2019. https://doi.org/10.1787/1d0bc92a-en

[13] L. D. Prasojo et al., "Teachers' burnout: A SEM analysis in an Asian context," Heliyon, vol. 6, no. 12, p. e03144, 2020. https://doi.org/10.1016/j.heliyon.2019.e03144

s

ICAIMT: Explainable AI for Learning in Higher Education

Meekah Ludba Abdul Hakeem
Dept. of Statistics and Business Analytics, College of Business UAE University
Al Ain-UAE
700050595@uaeu.ac.ae

Nouf Mohammed Hasan Alobeidli
Dept. of Statistics and Business Analytics, College of Business UAE University
Al Ain-UAE
201600181@uaeu.ac.ae

Dr Ananth Chiravuri Associate Prof
Dept. of Statistics and Business Analytics, College of Business UAE University
Al Ain-UAE
ananth.chiravuri@uaeu.ac.ae

Abstract— The objective of this study is to examine the impact of explainable AI (XAI) on student learning in higher education. Drawing on Social cognitive and Flow theories, we propose that XAI enhances trust, engagement, self-efficacy and flow experience. Findings from our survey indicate that trust in XAI significantly predicts engagement, self-efficacy and flow. Findings suggest that transparency in AI systems plays a critical role in shaping positive behavioural outcomes. Implications for higher educational institutions and AI developers are discussed.

Keywords—Explainable Artificial Intelligence (XAI), Trust, Self-Efficacy, Engagement, Flow Experience, Higher Educations.

I. INTRODUCTION

Artificial Intelligence (AI) and its related applications have rapidly gained prominence within higher education institutions (HEI's), fundamentally transforming how students access information, interact with educational resources, and engage in learning processes. The integration of Generative AI (GenAI) tools and intelligent learning applications such as ChatGPT, Co-Pilot, Gemini has enabled increasingly personalized, adaptive, and efficient learning experiences for students, while supporting instructors through automated feedback, content generation, and learning analytics [1]. As a result, AI-driven tools and systems are increasingly being embedded in teaching practices, assessment methods and pedagogical models.

Despite these advantages, the accelerated development and widespread adoption of AI-based tools in HEI's have raised significant concerns, particularly relating to transparency, accountability, and student trust [2]. Many AI systems operate using complex algorithms and machine learning models that are not easily interpretable by users. This so-called "black box" nature of AI presents a critical challenge in educational contexts, where understanding the rationale behind decisions, recommendations, or feedback is essential for meaningful learning and developing insight. When students and other users (e.g., instructors, teachers etc) are unable to comprehend how an AI system arrives at its outputs such as grading suggestions, learning recommendations, or academic guidance, they may question the system's reliability and fairness, leading to reduced trust and engagement. Thus, it is possible that a lack of transparency in AI systems can negatively influence students' learning experiences by diminishing confidence in AI-supported educational tools and potentially discouraging their effective use. Moreover, opaque decision-making processes may reinforce biases, limit opportunities for critical reflection, and undermine the pedagogical value of AI tools and technologies [3]. In environments that prioritise academic integrity, fairness, and learner autonomy, such limitations pose serious risks to educational outcomes and institutional credibility.

In response to these challenges, Explainable Artificial Intelligence (XAI) has emerged as a promising approach to enhance transparency and trust in AI-driven educational systems [4]. XAI focuses on making AI models more interpretable by providing clear, understandable explanations for their outputs and recommendations [5]. This enables students and educators to grasp the "how" and "why" behind AI generated decisions, thereby promoting more informed, confident and immersive learning experiences [6][7]. Consequently, the integration of explainability into AI-based educational tools represents a critical step forward in responsible and ethical AI adoption in higher education.

While XAI offers advantages over traditional black box systems, empirical research examining its educational impact remains limited. Most existing studies focus primarily on the adoption and usage of AI tools by students, whereas little is known about the specific effects of explainability on student learning outcomes [8] [9]. Accordingly, there is a need to systematically examine how XAI influences learning related concepts, forming the motivation for this study. Although many people are excited about AI as a technology including its tools in higher education, there is still insufficient understanding of whether and how XAI enhances students' learning experiences [10]. Most prior studies emphasize the technical side or focus on isolated outcomes, like trust or academic performance. Presently, very few studies have developed and empirically tested an integrated and inclusive model that examines multiple student psychological and behavioural dimensions of learning.

Although some studies have explored students' perceptions of AI in educational contexts, limited attention has been given

to how "explainability" affects interconnected learning variables such as trust, engagement, self-efficacy, and flow in a unified model. This is important because these variables are theoretically interrelated and collectively shape meaningful learning experiences. Therefore, understanding their interaction can help provide deeper insight into the mechanisms through which XAI may influence HE learning outcomes. Next, we present the literature review.

II. LITERATURE REVIEW

A. Explainable Artificial Intelligence (XAI) and Learning

Explainable AI refers to AI systems that are designed to make their decision making processes transparent and interpretable to human users. Normal "black-box" AI gives answers without explaining how it got them, but XAI shows the reasons, the data, and the steps used to make each choice. This transparency reduces information asymmetry between humans and algorithms, thereby allowing users to evaluate the logic behind the suggestions and/or recommendations.

In HEI's, this is very important because HEI students need to know not just *what* a concept means, but also develop a critical understanding as to *why* and *how* these concepts work. This is justified because Bloom's taxonomy (1956), indicates that students in HEI's need to focus more applying, analyzing, synthesizing and evaluating concepts rather than just knowledge and comprehension as done earlier [11]. Thus, explainability aligns with higher order cognitive processes required in HEI's where reasoning is critical.

Some scholars have explored XAI in institutions. They found 15 different meanings of XAI and 62 problems, grounded into seven areas: making things clear, being fair, computer parts, how people and computers work together, trust, rules, and other issues [12]. But their study noted that no one agreed to one meaning of XAI, which makes it confusing and problematic to get a consensus on. This conceptual fragmentation limits cumulate knowledge development. But even with this problem, most agree that XAI is very important for making AI easier to use and more trusted in learning. In particular, transparency has been associated with greater perceptions of reliability, accountability and fairness in algorithmic systems [13].

Some of them looked at different ways to make AI explain itself in institutions. For example, they examined tools like LIME and SHAP, which help explain any AI model, and simple models like decision trees that are easy to understand by themselves. They found that how well these tools work depends on where they are used, how much people know about technology, and what learning goals the teacher or student wants to reach. More importantly, prior research suggests that the effectiveness of explanations is contingent upon user characteristics such as knowledge, cognitive ability thereby implying that explainability must be context sensitive rather than purely technical [14]. Therefore, in HEI's, XAI should be designed and implemented to maximize its impact on trust, engagement and learning outcomes.

B. Social Cognitive Theory and AI-supported Learning

The Social Cognitive Theory (SCT) provides important theoretical foundations for understanding how students interact with technology-supported learning environments. According to the previous papers [15], a central construct within this theory is self-efficacy, defined as an individual's belief in their ability to perform tasks successfully. Students with stronger self-efficacy beliefs tend to demonstrate higher levels of effort, resilience, and participation in academic tasks [16]. In digital learning environments, technological systems can function as environmental factors that influence learners' behaviours and beliefs by providing feedback, guidance, and learning support [17].

The transparency in XAI Learning can strengthen the students' confidence by using AI tools effectively, thereby improving the perceived self-efficacy. In addition to this, the explainable systems can promote more active learner participation by enabling students to critically evaluate AI-generated suggestions and integrate them into their learning strategies. Prior research suggests that transparent and interpretable learning technologies can enhance learner motivation and engagement by increasing the perceived control and understanding of the learning process [18][19]. Therefore, from a social cognitive perspective, the explainable AI learning may positively influence students' self-efficacy and engagement in AI-supported learning environments.

C. Flow Theory and AI-supported Learning

The Flow theory [7] suggests that the flow occurs only when the learners perceive a balance between the challenge of a task and their own skills, with clear goals and immediate feedback. In educational contexts, the experience of flow has been associated with increased motivation, engagement, and improved learning performance. Research in the digital learning environments suggests that interactive technologies and well-designed learning systems can facilitate flow experience by providing structured guidance, continuous feedback, and meaningful interaction with learning tasks [20][21].

The transparency in the XAI enables the learners to focus more effectively on the task itself rather than questioning the system's recommendations. In addition to that, the trust in AI systems may further support the development of the flow experiences by allowing students to rely on the AI-supported guidance without any hesitation. Prior research indicates that the digital learning environments that provide clear feedback and interactive engagement can enhance learners' immersion and intrinsic motivation [22][23]. Therefore, explainable AI and trust in the AI systems may facilitate flow experiences by enabling sustained attention, deeper engagement, and more immersive learning interactions in AI-supported educational environments.

Based on the relevant theories, we argue and develop the following hypotheses. As discussed, XAI leads to more transparency, predictability and increased explainability, which in turn increases the reliance of the users on such systems, leading to greater trust in the AI systems. Therefore, we posit:

H1:Perceived explainable AI positively predicts students' trust in AI recommendations.

In addition, we hypothesize that perceived XAI could directly impact the three learning outcomes because of the underlying mechanisms of better structured guidance and explicit reasoning and test for the below:

H2a: Perceived explainable AI positively predicts students' self-efficacy in using AI for learning.

H2b: Perceived explainable AI positively predicts students' learning engagement.

H2c: Perceived explainable AI positively predicts students' flow experience during AI-supported learning.

Similarly, we posit greater trust will directly impact learning outcomes as well because when students trust XAI systems, they are more willing to rely on recommendations and strengthen their belief systems, and hypothesize:

H3a: Trust in explainable AI positively predicts students' self-efficacy.

H3b: Trust in explainable AI positively predicts students' learning engagement.

H3c: Trust in explainable AI positively predicts students' flow experience.

Finally, as indicated earlier, XAI provides transparency, which fosters trust by increasing perceptions of reliability and competence. Thus, we hypothesize trust to mediate the relationship between XAI and the learning outcomes and posit

H4a: Trust mediates the relationship between perceived explainable AI and students' self-efficacy.

H4b: Trust mediates the relationship between perceived explainable AI and students' learning engagement.

H4c: Trust mediates the relationship between perceived explainable AI and students' flow experience.

III. METHODOLOGY

In this section, we describe the methodology that was followed to conduct this study.

A. Research Design

This study adopted a quantitative analysis using a survey that included a questionnaire . This approach was appropriate for looking at the relationships between perceived explainability of AI (independent variable) and other variables measuring learning (dependent variables). These dependent variables include Trust (mediator variable), Engagement, Self-Efficacy, Flow Experience. We used a survey method as it enabled efficient collection of data from many students, thereby supporting the statistical examination of the proposed relationships between XAI, trust and the dependent variables. We administered the survey online to facilitate faster data collection.

B. Data Collection

The survey was administered online using Google forms, ensuring accessibility and enabling efficient data collection. The survey link was distributed to students in HEI's through multiple channels such as direct institutional email, announcements in learning management systems (Blackboard), social media platforms used by UAEU students, and in-class announcements by cooperative faculty members. The survey began with an informed consent statement explaining the research purpose and procedures to respondents including voluntary participation, confidentiality protections, right to withdraw, and researcher contact information. Participants indicated consent by proceeding with the survey after reviewing this information. Data collection occurred over four weeks during the academic semester, avoiding examination periods to ensure adequate response rates and minimize stress-related response bias. Till date, a total of 79 usable responses were obtained. We analyzed the data using regression as it was suitable to test for direct effects and mediation pathways.

IV. INITIAL FINDINGS

Our initial findings revealed that perceived explainability of AI significantly predicts trust in AI recommendations ($\beta = .725$, $p < .001$), supporting H1. Our findings also indicated that XAI directly predicts self-efficacy and flow. More importantly, these effects remained significant even after accounting for trust, which suggests that explainability may enhance students' sense of competence and immersion through clarity, guidance, and comprehensibility of explanations (i.e., the explanation itself reduces uncertainty and increases perceived capability), rather than via trust.

We also found that trust significantly predicted student engagement in their studies, but not their confidence in using AI supported technologies, and their ability to remain more deeply focused on their learning [24]. XAI increases engagement because, as seen in H1, it increases trust. Students invest time, stay active, and remain focused when they perceive recommendations as trustworthy.

Finally, our findings indicate that trust mediates the relationship between perceived XAI and engagement. However, trust did not function as a universal psychological mechanism across all outcomes, as no mediation effects were observed for self-efficacy or flow. This pattern suggests that trust plays a selective role in translating explainability into behavioural engagement, whereas cognitive and experiential outcomes may stem more directly from the clarity provided by explanations themselves.

V. CONCLUSION AND LIMITATIONS

Our findings are strategically important for universities and companies that make educational technology. They suggest that XAI is not merely a nice feature but a foundational feature that determines whether AI tools are adopted, trusted and effectively utilized in HEI's.

The improvement of student learning appears to be facilitated when students exhibit higher levels of trust in the learning platform that they use. Our study indicates that organizations developing AI based educational platforms should prioritize transparency and explainability to impact the learning outcomes and enhance user trust, which in turn may strengthen user engagement. Platform evelopers shoul

consider trust building mechanisms as a core design principle. Future education in AI depends not only on technological advancement but also on the integration of meaningful and context driven explanations.

When a student understands the AI's advice, their perceived self-efficacy in using AI systems goes up leading to improved learning experiences which then could create a state of flow. Transparency functions as a mechanism to build trust that enables higher engagement, greater confidence and deeper learning states [25].

A limitation of our study is that it mainly focuses on four areas: Trust, Engagement, Self-efficacy, Flow experience. Future studies can investigate other dimensions across different countries to assess cultural variability and enhance external validity. To conclude, our study indicates that XAI and students' trust in AI systems plays a central mediating role in shaping students' learning outcomes in different ways thereby supporting a more meaningful and immersive learning experience.

REFERENCES

[1] Schepman, A., & Rodway, P. (2022). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human-Computer Interaction, 39*(13), 2724-2741.

[2] European Commission. (2019). *Ethics guidelines for trustworthy AI*. High-Level Expert Group on Artificial Intelligence, European Union.

[3] Dawson, P., Tai, J., Bearman, M., Boud, D., & Ajjawi, R. (2023). The development and validation of the assessment engagement scale. *Frontiers in Psychology, 14*, 1136878.

[4] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, 39.

[5] Gunasekara, S., & Saarela, M. (2025). Explainable AI in education: Techniques and qualitative assessment. *Applied Sciences, 15*(3), 1239.

[6] Csikszentmihalyi, M. (1975). Beyond boredom and anxiety. Jossey-Bass.

[7] Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. Harper & Row.

[8] Pearce, J. M., Ainley, M., & Howard, S. (2005). The ebb and flow of online learning. *Computers in Human Behavior, 21*(5), 745-771.

[9] Wang, Y. Y., & Chuang, Y. W. (2024). Artificial intelligence self-efficacy: Scale development and validation. *Education and Information Technologies, 29*(4), 4785-4808.

[10] Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.*

[11] B. S. (1956b) Taxonomy of Educational Objectives; the Classification of Educational Goals, by a Committee of College and University Examiners., David McKay, New York, NY.

[12] Altukhi, A., & Pradhan, S. (2024). Systematic literature review: Explainable AI definitions and challenges in education. *arXiv preprint* arXiv:2504.02910.

[13] Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies*, 146, 102551.

[14] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

[15] Arastaman, G. (2014). Validity and reliability of student engagement scale. *Bartin University Journal of Faculty of Education, 3*(2), 390-403.

[16] Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

[17] Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

[18] Wang, Y., et al. (2023). Artificial intelligence in education and learner self-efficacy: Implications for engagement and learning outcomes. Computers in Human Behavior, 139, 107672. DOI: 10.1016/j.chb.2023.107672

[19] DiBenedetto, M. K., & Zimmerman, B. J. (2018). Self-regulation and social cognitive theory in education. Educational Psychologist, 53(1), 35–50. DOI: 10.1080/00461520.2018.1465313

[20] Rodríguez-Ardura, I., & Meseguer-Artola, A. (2017). Flow in e-learning environments: Direct and indirect effects on learning outcomes. Computers in Human Behavior, 72, 699–706. DOI: 10.1016/j.chb.2017.01.039

[21] Buil, I., Catalán, S., & Martínez, E. (2020). Understanding flow in digital learning environments. Computers & Education, 149, 103821. DOI: 10.1016/j.compedu.2020.103821

[22] Kiili, K., et al. (2021). Flow experience and engagement in digital learning environments. Educational Technology Research and Development, 69(1), 361–385. DOI: 10.1007/s11423-020-09864-8

[23] Engeser, S., & Rheinberg, F. (2018). Flow, performance and motivation. Motivation and Emotion, 42(5), 1–15. DOI: 10.1007/s11031-018-9696-2

[24] Arastaman, G. (2014). Validity and reliability of student engagement scale. *Bartin University Journal of Faculty of Education, 3*(2), 390-403.

[25] Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215

Tacit Knowledge Seeking in the Age of AI: Implications and Safeguards

Dr. Mamie Yvette Griffin
Faculty of Business
Higher Colleges of Technology, Abu Dhabi Colleges
Abu Dhabi, UAE
mgriffin@hct.ac.ae

Abstract—This paper examines how the increasing use of artificial intelligence (AI) may influence employees' knowledge-seeking behavior and its implications for tacit knowledge development within organizations. A structured literature review of prior research centered on tacit knowledge, artificial intelligence in the workplace, and knowledge-seeking behavior was conducted. A thematic analysis of relevant literature suggests a dual effect: AI may reduce interpersonal interaction and weaken informal learning processes, while also enabling greater collaboration by reducing time spent on routine tasks. These contrasting effects highlight a tension between efficiency and social learning in AI-enabled environments.

This paper addresses the limited research on how AI influences employees' knowledge-seeking behavior and offers directions for preserving tacit knowledge in AI-enabled work environments.

Keywords—Tacit Knowledge; AI; Knowledge-Seeking Behavior; Organizational Learning

I. INTRODUCTION

Artificial Intelligence (AI) is rapidly reshaping the way many employees complete organizational tasks. From ChatGPT to more advanced tools, employees can quickly access data for various purposes including information retrieval, problem solving, and decision making. Employees can obtain answers, generate options, and complete tasks with less delay than in many traditional knowledge systems. From an organizational standpoint, these capabilities promise gains in efficiency and productivity. Supporting this view, a 2023 report found that generative AI increased productivity in knowledge-intensive tasks by 40 to 60 percent [13].

However, the growing use of AI raises concerns beyond productivity measures. Traditionally, employees relied on experienced colleagues for assistance. For example, new employees would naturally seek the guidance of senior employees to understand a workplace situation. Even experienced employees might consult knowledgeable coworkers when facing unfamiliar issues, complex decisions, or situations requiring specialized expertise. Such interactions did more than simply provide answers; they enabled the transfer of tacit knowledge.

Tacit knowledge is experience-based and rooted in experience, judgment, and situational understanding. This form of knowledge is difficult to articulate and formalize [3], yet it remains a critical source of organizational learning, capability and competitive advantage. Tacit knowledge provides employees practical insights and contextual understanding of workplace issues.

Beyond understanding what tacit knowledge is, it is equally important to understand how it is transferred within organizations. Traditionally, tacit knowledge is passed from one colleague to another through collaborative practices and interpersonal processes including social interactions, observation, mentoring, and informal conversations. It involves navigating social networks to identify "who knows what" in an organization. The process of seeking tacit knowledge contributes not only to learning, but also trust building, team development, and the gradual development of expertise.

As AI becomes more prominent in organizations, opportunities for interpersonal learning may decline. The same employees who would normally ask a senior colleague for advice might increasingly rely on an AI assistant or chatbot for immediate support. While this shift may produce efficient results, it raises an important question: what happens to tacit knowledge if employees stop asking one another for help?

Although existing literature examines AI adoption and knowledge management from multiple perspectives, little attention has been given to how AI may influence tacit knowledge-seeking behavior. This issue is developed further in the following problem statement.

II. PROBLEM STATEMENT

With the integration of AI to enhance organizational efficiency and productivity, a potential paradox emerges. The same technology intended to improve knowledge management could also weaken the foundation of tacit knowledge – the informal employee to employee networks through which tacit knowledge is accessed and transferred. Gradually, the impact could lead to reduced interpersonal learning, experiential knowledge sharing, and the subtle erosion of valuable organizational knowledge.

To better understand this concern, further research is needed on the risks AI may pose to tacit knowledge processes. While recent studies explore the intersection of these two topics from various perspectives including AI adoption; knowledge capture, processing, and maintenance; and the benefits of intelligent systems, there remains substantial room

to explore how AI affects employees' knowledge-seeking behaviors and the consequences to tacit knowledge.

This gap is significant because weakened informal learning networks could undermine tacit knowledge exchange which is essential for organizational learning and can ultimately reduce an organization's resources, capabilities, and competitive advantage.

III. RESEARCH QUESTIONS

This study is guided by following research questions:

1. How does the use of artificial intelligence influence tacit knowledge-seeking behavior in organizations?
2. What are the implications of reliance on artificial intelligence for tacit knowledge transfer within organizational contexts?
3. How can organizations preserve tacit knowledge-seeking in AI enabled environments?

IV. LITERATURE REVIEW

This study draws on two interconnected streams of literature: tacit knowledge in organizations and artificial intelligence in knowledge work.

The literature on tacit knowledge emphasizes its experiential, context-specific, and difficult-to-codify nature [2] [3]. Tacit knowledge is developed through practical experience and is primarily transferred through informal communication and social processes such as interpersonal connections, mentoring, observation, and collaboration [1]. Within knowledge management research, it is widely recognized as a key driver of organizational learning, innovation, and sustained competitive advantage. As noted by some researchers, tacit knowledge may be an organization's most significant resource [4].

The second stream - artificial intelligence in organizations - is addressed in current literature with some significance. Although research on AI in knowledge management is still developing, several studies highlight AI's increasing role in supporting knowledge work [5]. Evidence suggests that AI has been adopted by various entities to improve knowledge acquisition, provide solutions to problems, optimize solution systems, and for modeling purposes [6].

Examined independently, both research streams provide valuable insights for organizations. Existing studies have begun to explore the impact of AI on knowledge management processes [7] including its associated challenges [8] and benefits [9]. Other studies examine how to successfully integrate AI with knowledge management and other organizational factors [10] [11].

While the combination of these topics continues to be explored from various angles, an important gap remains. The tacit knowledge literature assumes that employees seek knowledge through interpersonal interaction, whereas the AI literature often focuses on a shift toward tool-based knowledge access. However, limited research examines how potential shifts in tacit knowledge-seeking behavior—from reduced reliance on interpersonal interactions to consulting AI—may impact the development and transfer of tacit knowledge. This shift may alter the processes that sustain tacit knowledge within organizations as well as intended outcomes. The present study seeks to address this gap.

V. CONTRIBUTIONS

This paper extends the conversation on artificial intelligence and knowledge management in two important ways. First, it explores an emerging tension between AI adoption and tacit knowledge processes by examining how AI reshapes employees knowledge-seeking behavior. This shifts the lens from the technical benefits of AI adoption to a behavioral perspective that has received less emphasis, highlighting how changes in knowledge-seeking practices may influence tacit knowledge development.

Secondly, the paper introduces a conceptual explanation of how employees' increased reliance on AI in their everyday work setting may reduce opportunities for interpersonal consultation, thus impacting tacit knowledge transfer. AI makes it easier to access information, but it may reduce the need for colleagues to interact. This shifts the conversation from short-term productivity gains to long-term capability development.

Furthermore, it provides practical recommendations for management to successfully implement AI without losing the human interactions critical to tacit knowledge development and transfer.

VI. METHODOLOGY

The study adopts a qualitative conceptual design based on a structured review of existing literature using Denyer and Tranfield's five-step systematic review approach [14]. This framework provides a transparent and structured process for synthesizing prior research. It consists of five stages: (1) question formulation, (2) locating studies, (3) study selection and evaluation, (4) analysis and synthesis, and (5) reporting the results.

The key question revolved around developing a clearer understanding of how AI may influence tacit knowledge-seeking behavior in organizational settings. A collection of academic articles was obtained using various databases including Scopus, Web of Science, and Google Scholar using a systematic process. The literature search focused on three key topic areas:

1. Tacit Knowledge: foundational and contemporary studies that explore the meaning, value, and transfer of tacit knowledge within organizations.
2. AI in the Workplace: studies that examine how AI influences the actions and behaviors of employees.

3. Employees Knowledge-Seeking Behavior: Research exploring how employees seek, share, and apply knowledge in organizational settings.

Articles were selected for deeper review if they addressed one or more of the following criteria:

- Peer reviewed journal articles
- Publications within the past 10 years, with exceptions for earlier seminal publications.
- Studies examining tacit knowledge, knowledge management, and AI in the workplace.

At present, a total of 17 peer-reviewed articles have been retained for inclusion in the review and analyzed using thematic analysis to identify recurring patterns, key concepts, and emerging themes related to AI, tacit knowledge, and employee knowledge-seeking behavior.

A conceptual review is appropriate for this topic because organizational use of AI is still evolving, and empirical evidence regarding its long-term effects on tacit knowledge remains limited. Therefore, this approach provides a foundation for future survey, case-study, and longitudinal research.

VII. PRELIMINARY FINDINGS

Drawing on thematic analysis of the reviewed literature, the following propositions are proposed:

P1: Reliance on AI systems rather than colleagues for knowledge - seeking reduces interpersonal social interactions, which in turn weakens the development and exchange of tacit knowledge over time.

P2: Reliance on AI generated solutions without colleague engagement in the problem-solving process reduces experiential learning within teams and weakens the development of tacit knowledge over time.

P3: AI use increases the frequency of employee social interactions by reducing time spent on routine tasks, thereby enabling greater opportunities for higher order thinking and collaborative discussion that enhances tacit knowledge over time.

P4: Managerial policies and organizational practices influence how the tension between AI-driven efficiency and socially embedded tacit knowledge processes is resolved.

VIII. IMPLICATIONS

The outcomes of this study present several important implications that aid in understanding the impact of artificial intelligence on employee's knowledge-seeking behavior. The proposed propositions suggest that AI's influence is complex with pathways that potentially enhance and erode tacit knowledge access and development.

The propositions (P1, P2) suggest that increased reliance on AI may reduce interpersonal interaction and limit engagement in collaborative problem solving. As employees shift their knowledge-seeking behavior from consulting colleagues to relying on AI prompts, opportunities for social interactions may decline. This shift may weaken knowledge ties between employees, reducing exposure to experiential insights and informal learning that typically emerge from social interactions. In other words, AI may help employees obtain solutions, but the underlying reasoning processes that typically develop from interacting with experienced colleagues may be diminished, resulting in the decoupling of problem solving from learning.

Furthermore, as employees increasingly use AI instead of interacting with colleagues the strength, quality and role of informal organizational networks in facilitating knowledge sharing may lessen. Even if such ties remain intact, reduced engagement limits the depth of interactions. Consequently, opportunities for collaboration and shared understanding may decline, potentially affecting how knowledge is disseminated across the organization.

While P1 and P2 highlight the constraining effects of AI on knowledge-seeking behavior, P3 suggests a contrasting outlook whereby AI creates opportunities for increased employee interactions by reducing the time typically spent on routine tasks such as information retrieval, report generation, or basic data analysis. Freed from time-consuming mundane tasks, employees have greater availability for social interactions including informal exchanges such as conversations during coffee breaks. These interactions strengthen personal bonds that facilitate future sharing and reinforce the social processes underlying tacit knowledge development.

Collectively, these implications suggest that while AI can improve efficiency in knowledge access, it may also unintentionally disrupt the social processes that define tacit knowledge development. Therefore management must ensure that AI adoption is accompanied by practices and policies aimed at sustaining interpersonal interaction and the continual development of tacit knowledge [P4]. To this end, the study provides practical recommendations for organizations to preserve tacit knowledge - seeking in AI-enabled workspaces. Some key recommendations are:

Organizations must recognize that far from simply improving efficiency, AI changes how people learn and interact.

Established policies for AI usage should emphasize the use of AI to support work while keeping human interaction intact.

Encourage interaction through active mentoring and coaching programs.

Formalize informal conversations with organizational practices and events that drive informal knowledge exchange.

Promote AI-aware knowledge training to help employees develop judgement in seeking AI support or human support.

IX. CONCLUSION AND FUTURE RESEARCH

This study highlights an emerging tension between the increasing use of artificial intelligence and the social

processes that facilitate tacit knowledge development within organizations. By examining how AI may shift employees' knowledge-seeking behavior away from interpersonal interaction, the paper suggests that AI adoption may have unintended consequences for informal learning and tacit knowledge development. Given the conceptual nature of this research, future empirical studies are needed to examine how AI influences knowledge-seeking behavior in practice. Additionally, longitudinal research is required to understand AI's long-term effects on knowledge-seeking behavior and tacit knowledge development.

REFERENCES

[1] Ali, H.A.A., Elanain, H.M.A., and Ajmal, M.M. (2016), "Knowledge sharing-behaviour as a mediator of the relationship between organizational justice and organizational performance in the UAE", *International Journal of Applied Management Science*, Vol. 8, No. 4, pp. 290-312.

[2] Nonaka, I. (1991), "The Knowledge-Creating Company", *Harvard Business Review*, Vol. 69, No. 6, pp. 96-104.

[3] Polanyi, M. (1969), "The logic of tacit inference", in Grene, M. (Ed.), *Knowing and Being*, Routledge & Kegan Paul, London.

[4] McAdam, R., Mason, B., and McCrory, J. (2007), "Exploring the Sdichotomies within the tacit knowledge literature: towards a process of tacit knowing in organizations", *Journal of Knowledge Management*, Vol. 11, No. 2, pp. 43-59.

[5] Alhashmi, S.F.S., Salloum, S.A., and Abdallah, S. (2019), "Critical success factors for implementing artificial intelligence (AI) projects in Dubai Government United Arab Emirates (UAE) health sector: Applying the extended technology acceptance model (TAM)", in *International Conference on Advanced Intelligent Systems and Informatics*, Springer, Berlin/Heidelberg, Germany.

[6] Taherdoost, H., and Madanchian, M. (2023), "Artificial intelligence and knowledge management: Impacts, benefits, and implementation", *Computers*, Vol. 12, No. 4, p. 72.

[7] Nakash, M., and Bolisani, E. (2025), "The transformative impact of AI on knowledge management processes", *Business Process Management Journal*, Vol. 31, No. 8, pp. 124-147.

[8] Njiru, D.K., Mugo, D.M., and Musyoka, F.M. (2025), "Ethical considerations in AI-based user profiling for knowledge Smanagement: A critical review", *Telematics and Informatics Reports*, Vol. 18, p. 100205.

[9] Arakpogun, E.O., Elsahn, Z., Olan, F., and Elsahn, F. (2021), "Artificial Intelligence in Africa: Challenges and Opportunities", in *The Fourth Industrial Revolution: Implementation of Artificial Intelligence for Growing Business Success*, pp. 375-388.

[10] Olan, F., Arakpogun, E.O., Suklan, J., Nakpodia, F., Damij, N., and Jayawickrama, U. (2022), "Artificial intelligence and knowledge sharing: Contributing factors to organizational performance", *Journal of Business Research*, Vol. 145, pp. 605-615.

[11] Sanzogni, L., Guzman, G., and Busch, P. (2017), "Artificial intelligence and knowledge management: questioning the tacit dimension", *Prometheus*, Vol. 35, No. 1, pp. 37-56.

[12] Retkowsky, J., Hafermalz, E., and Huysman, M. (2024), "Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects", *Business Horizons*, Vol. 67, No. 5, pp. 511-523.

[13] Yan, J., Husted, K., and Fath, B. (2026), "Transforming organizational knowledge creation through artificial intelligence: a systematic review of the emergent literature", *VINE Journal of Information and Knowledge Management Systems*, Vol. 56, No. 2, pp. 522-540.s

[14] Denyer, D., and Tranfield, D. (2009), "Producing a systematic review", in Buchanan, D.A. and Bryman, A. (Eds), *The Sage Handbook of Organizational Research Methods*, Sage Publications Ltd, pp. 671-689.

Rethinking AI-Driven Customer Loyalty in Retail

Shamma Alsaedi
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
adsm-215909@adsm.ac.ae

Ishtiaq Rasool Khan
Business Analytics Program
Abu Dhabi School of Management
Abu Dhabi, United Arab Emirates
i.khan@adsm.ac.ae
ORCID: 0000-0002-3887-9052

***Abstract*—Artificial intelligence (AI) is being increasingly used in retail to automate service delivery and improve personalization, and one of the goals of using AI is to strengthen customer loyalty. In this paper, we report findings from a small exploratory study in the United Arab Emirates (UAE) that examines how retail consumers relate AI-enabled experiences to their loyalty intentions. The study proposes and pilots a structured survey instrument capturing AI awareness, perceived personalization, perceived value, service convenience, trust, perceived intrusiveness, and loyalty in UAE retail settings. Survey data were collected from 25 respondents and used to assess directional associations between these constructs. Contrary to widely held expectations, the results show that more positive evaluations of AI-enabled retail services were not associated with higher loyalty; in several cases, the observed relationships were weak, non-significant, or negative. Trust levels were also generally modest, suggesting that functional improvements enabled by AI may not translate into stronger relational outcomes in consumer-retailer relationships. While the small sample size limits generalization and the study should be interpreted as an early pilot, the findings raise important questions about the assumed loyalty benefits of AI adoption in retail. The paper contributes early evidence from a UAE context and highlights trust and human-centered service design as central considerations for future large-scale research.**

Keywords—Artificial intelligence in retail; Customer loyalty; Trust and privacy; Service automation; Human-centric AI

I. INTRODUCTION

Artificial intelligence (AI) is now part of the competitive strategies adopted by most retailers. It supports personalization, automates parts of customer service, and enables data-driven decision making at scale. AI recommendation systems, chatbots, and predictive analytics are widely used to improve efficiency and customer experience, and retailers often assume that these improvements will naturally lead to stronger customer loyalty (Shankar, 2018; Davenport et al., 2020). In other words, AI is increasingly treated not only as a tool for operations, but also as a way to strengthen long-term customer relationships. In addition to personalization and service automation, AI is increasingly used for customer sentiment and insight generation from digital interactions, further strengthening its role in data-driven customer understanding (Bibi et al., 2025).

Yet the link between AI adoption and loyalty is not as clear as industry narratives suggest. AI-driven personalization and automation can improve convenience and perceived value, but loyalty is not purely functional. It is also shaped by relational and emotional factors that technology may not fully replicate (Huang & Rust, 2018; Prentice et al., 2020). Consumers may appreciate speed and personalization, while still feeling uneasy about how their data are used, how transparent AI decisions are, or whether the shopping experience is becoming too impersonal. Studies increasingly highlight the role of trust, privacy concerns, and the perceived loss of human interaction in shaping consumer responses to AI-enabled services (Van der Aa et al., 2021; Puntoni et al., 2021). This creates a broader tension in AI-driven retail, often framed as a personalization–privacy and automation–humanity trade-off (Rosado-Pulido et al., 2023).

Despite growing research in this area, two issues remain underexplored. First, much of the literature focuses on adoption, satisfaction, or short-term service outcomes, while the assumed translation of AI improvements into customer loyalty is examined less consistently. Second, trust-related concerns such as perceived intrusiveness are frequently discussed as important drivers of acceptance, but they are not always measured alongside loyalty outcomes in empirical studies. These gaps matter because AI may raise service performance while failing to strengthen the relational foundations that loyalty depends on.

These questions are especially relevant in rapidly digitalizing retail environments such as the United Arab Emirates (UAE). The UAE retail sector has high levels of technological investment and strong competitive pressure, alongside a culturally diverse consumer base. At the same time, there is increasing public and regulatory attention to data protection and responsible AI use. While AI adoption in the region continues to accelerate, there remains limited consumer-centered evidence on how AI-enabled retail experiences relate to loyalty in practice.

Against this background, the present study reports an exploratory pilot investigation of AI-driven retail experiences and customer loyalty in the UAE. The study proposes and pilots a structured survey instrument capturing AI awareness, perceived personalization, perceived value, service convenience, trust, perceived intrusiveness, and loyalty. Using descriptive and exploratory analysis, the study examines whether the commonly assumed positive relationship between AI-enabled retail services and loyalty is reflected in customer perceptions, and whether trust-related tensions emerge as a potential explanation. Given the pilot nature of the study and its small sample size, the results are not intended to support causal claims. Instead, the aim is to provide early empirical

signals that can inform future large-scale studies and help refine how loyalty is conceptualized in AI-enabled retail.

The contribution of this conference paper is threefold. First, it provides early empirical insight into AI–loyalty dynamics within an underexplored regional context. Second, it places trust and perceived intrusiveness at the center of the analysis of consumer responses to AI-enabled retail services. Third, by reporting counterintuitive pilot findings, the study contributes to ongoing debates on the limits of AI-driven personalization and automation in fostering sustainable customer loyalty, while also motivating future theory-refining research.

II. RELATED WORK

A. AI-Driven Personalization and Perceived Value

A substantial body of literature positions AI-driven personalization as a key mechanism through which retailers enhance customer experience and loyalty. Studies consistently report that AI-enabled recommendation systems and data-driven personalization can improve perceived relevance, efficiency, and convenience of retail interactions, thereby strengthening engagement and satisfaction (Shankar, 2018; Davenport et al., 2020). From this perspective, personalization is treated as a value-creation tool that aligns offerings with individual customer preferences at scale.

However, empirical evidence also highlights important limitations to this assumption. Prentice et al. (2020) demonstrate that personalization contributes to engagement primarily when it is perceived as contextually appropriate and complemented by high-quality human service. Similarly, Rosado-Pulido et al. (2023) identify a personalization–privacy paradox, whereby increased personalization enhances perceived value while simultaneously intensifying privacy concerns that erode trust and loyalty intentions. These findings suggest that personalization alone is insufficient to guarantee loyalty and may become counterproductive when perceived as intrusive or misaligned with consumer expectations.

Despite these insights, much of the existing literature continues to treat personalization as a predominantly positive driver of loyalty, often assuming linear relationships between personalization intensity and relational outcomes. There is limited empirical work examining situations in which perceived value gains from AI personalization coexist with declining loyalty, particularly in culturally diverse and privacy-sensitive retail environments.

B. AI-Based Service Automation and Customer Experience

Beyond personalization, AI-driven service automation, particularly through chatbots and virtual assistants, has been widely studied as a means of improving service quality and operational efficiency. Research in hospitality and retail contexts indicates that AI-based service tools can enhance responsiveness, availability, and process convenience, contributing positively to customer satisfaction and usage intentions (Pillai & Sivathanu, 2020; Ameen et al., 2021).

At the same time, several studies caution that service automation introduces risks related to empathy deficits, inflexibility, and service failure. Pillai and Sivathanu (2020) show that while perceived usefulness and ease of use support adoption, chatbot failures or lack of emotional sensitivity can undermine trust and negatively affect loyalty intentions. Huang and Rust (2018) further argue that AI excels at mechanical and analytical service tasks but lacks emotional intelligence, which remains critical for relational loyalty formation. As a result, fully automated service encounters may weaken emotional bonds even when efficiency improves.

The literature thus reflects a growing consensus that hybrid service models combining AI efficiency with human intervention are more likely to sustain long-term customer relationships. However, empirical studies rarely investigate whether customers who rate AI-based service quality highly also exhibit stronger loyalty, or whether positive service evaluations may coexist with relational disengagement.

C. Trust, Privacy, and Ethical Concerns in AI-Enabled Retail

Trust is repeatedly identified as a central mediating mechanism in AI-driven customer experiences. Van der Aa et al. (2021) provide empirical evidence that trust mediates the relationship between AI capabilities and loyalty outcomes, emphasizing that efficiency and personalization gains do not translate into loyalty when trust is compromised. Similarly, Puntoni et al. (2021) highlight consumer concerns related to autonomy, control, transparency, and data usage, which can reduce trust and loyalty intentions despite improvements in convenience.

Retail-focused studies reinforce this view by demonstrating that perceived risk and lack of transparency can offset the benefits of smart retail technologies (Ameen et al., 2021). Davenport et al. (2020) further argue that ethical data practices and transparency are strategic prerequisites for sustainable AI-driven marketing outcomes. Collectively, these studies suggest that trust is not a secondary consideration but a foundational condition for AI-enabled loyalty.

Nevertheless, while trust is frequently theorized as a mediator, empirical findings on its role remain mixed, and many studies assume trust functions uniformly across contexts. There is limited evidence on how trust operates in emerging or rapidly digitalizing markets, where regulatory frameworks, cultural norms, and consumer expectations regarding data usage may differ substantially from Western contexts.

D. Gaps and Positioning of the Current Study

Across the literature, three dominant assumptions are evident. First, AI-driven personalization and service automation are generally expected to enhance loyalty through improved value and convenience. Second, trust is acknowledged as a critical mediator, yet often treated as a stable construct rather than a fragile or contested one. Third,

most empirical studies focus on mature markets, with limited attention to region-specific dynamics.

This study positions itself at the intersection of these gaps by empirically examining AI-enabled loyalty relationships in the UAE retail context through an exploratory pilot approach. Rather than assuming positive linear effects, the study allows for the possibility of counterintuitive associations between AI-related constructs and loyalty outcomes. By doing so, it responds to calls in the literature to critically examine the limits of AI-driven personalization and automation, particularly in environments where trust, privacy, and human-centric service expectations are salient (Huang & Rust, 2018; Rosado-Pulido et al., 2023).

III. METHODS AND DATA

A. Study design

This study uses a quantitative, cross-sectional exploratory design to examine how consumers in the UAE perceive AI-enabled retail services and how these perceptions relate to customer loyalty. The work is framed as a pilot study. Its purpose is to surface early directional patterns, assess how the constructs behave in a real survey setting, and identify potential tensions (for example around trust and privacy) rather than to test causal relationships or produce generalizable estimates. A survey was used to capture self-reported perceptions of AI-enabled retail interactions at a single point in time, which fits the study's exploratory scope.

B. Sample and data collection

Data were collected through a structured online questionnaire administered to adults (18+) residing in the UAE who reported recent exposure to AI-enabled retail services (such as recommendations, automated support, or AI-assisted service features). Due to practical constraints related to access, time, and resources, recruitment followed a non-probability approach combining purposive and snowball sampling. The purposive element ensured relevance (respondents had experience with AI-enabled retail), while snowball sharing helped reach additional participants within similar consumer networks.

The final pilot dataset includes 25 usable responses. Given the sampling approach and small sample size, results are treated as exploratory and indicative. Demographic variables (including age, gender, and shopping frequency) are reported descriptively to provide context, not to claim representativeness of the wider UAE retail population.

Ethical procedures were incorporated into the questionnaire. Participants were informed about the purpose of the study, that participation was voluntary, and that responses were anonymous. No personally identifiable information was collected.

C. Instrument and measures

The survey instrument consisted of 20 items organized around the study constructs. Perceptual items were measured using a five-point Likert scale (1 = Strongly disagree to 5 = Strongly agree), while demographic items used categorical response options. The questionnaire captured the following constructs:

- **AI awareness:** familiarity with AI features used by retailers
- **Perceived personalization and perceived value:** perceived relevance of recommendations and value of offers
- **Service convenience and AI-enabled service quality:** perceived ease and usefulness of AI-supported retail services (including automated support where applicable)
- **Trust and privacy concern:** confidence in retailers' data handling and perceived privacy-related risk
- **Customer loyalty:** repurchase intention and advocacy-related intentions

Where constructs were measured using multiple items, composite scores were created by averaging the relevant items. Items reflecting privacy concern were reverse-coded where needed so that higher composite values consistently reflected more positive evaluations (for example higher trust or lower perceived risk). The instrument was developed by mapping items to the study focus and was reviewed for clarity prior to distribution.

D. Data analysis approach

Analysis was conducted in SPSS using an exploratory workflow. First, the data were screened and summarized using descriptive statistics to profile the sample and examine the distributions and central tendencies of key constructs. Internal consistency of multi-item composites was assessed using Cronbach's alpha, with the recognition that reliability estimates can fluctuate substantially in small pilot samples.

Bivariate correlations were then used to examine directional associations between AI-related constructs and customer loyalty. Multiple linear regression was conducted in an exploratory manner to examine the combined association of AI awareness, perceived value, service quality, and trust with loyalty outcomes. Simple group comparisons (t-tests) and descriptive comparisons across age categories were also used to explore whether patterns differed across participant subgroups.

Given the pilot nature of the study, results are interpreted cautiously. The emphasis is placed on directionality, construct behavior, and potential tensions rather than on statistical confirmation. The intent is to inform instrument refinement and guide the design of future larger-scale studies. Given the small pilot sample, all statistical analyses are used strictly for exploratory and descriptive purposes. In particular, the correlation and regression results are retained only as preliminary diagnostic checks to identify possible directional patterns and measurement issues, not as confirmatory evidence of stable construct-level relationships.

IV. RESULTS

A. Sample Characteristics

The pilot sample consists of 25 UAE retail consumers with recent exposure to AI-enabled retail services. The age distribution is concentrated in younger and middle-aged groups, with 28% aged 18–25, 36% aged 26–35, and 20% aged 36–45, while smaller proportions fall in the 46–55 and 55+ categories. The sample includes 56% male and 44% female respondents. In terms of nationality, the sample reflects the UAE's demographic diversity, comprising Arab expatriates (48%), Asian expatriates (32%), UAE nationals (12%), and Western expatriates (8%). Most respondents reported frequent retail engagement, with 52% shopping monthly and 24% shopping weekly.

These characteristics indicate that the sample represents active retail consumers with varied demographic backgrounds, while remaining limited in size and representativeness due to the pilot design.

B. Descriptive Statistics of Key Constructs

Descriptive analysis shows high awareness of AI in retail contexts (mean = 4.28, SD = 0.84), suggesting that respondents are generally familiar with AI-driven features. Perceptions of personalization and perceived value are moderate, with a personalization mean of 3.32 (SD = 0.85) and offer value mean of 3.76 (SD = 0.93).

Perceived service convenience scores relatively high (mean = 4.16, SD = 0.80), whereas chatbot satisfaction is lower and more variable (mean = 3.20, SD = 1.12), indicating mixed evaluations of AI-based customer service interactions.

Trust-related measures reveal comparatively lower scores. Data trust has a mean of 2.88 (SD = 1.24), while privacy concern is moderate (mean = 3.44, SD = 1.00). After reverse coding privacy concern, the trust composite remains low (mean = 2.72, SD = 0.77), suggesting persistent consumer unease regarding data usage.

Customer loyalty indicators are moderate, with repurchase intention (mean = 3.40, SD = 1.00), brand advocacy (mean = 3.72, SD = 0.89), and an overall loyalty composite mean of 3.56 (SD = 0.67).

C. Reliability Assessment

Reliability analysis produced negative Cronbach's alpha values for both the Trust ($\alpha = -0.165$) and Loyalty ($\alpha = -0.019$) composites. These results indicate substantial measurement instability, likely attributable to the very small sample size, reverse-coded items, and potential multidimensionality of the constructs. As such, reliability coefficients are not interpreted as evidence of scale quality but rather as diagnostic indicators highlighting the need for scale refinement and validation in future studies.

D. Correlation Analysis

Correlation analysis reveals several counterintuitive associations. Service quality is significantly and negatively correlated with loyalty ($r = -0.489$, $p = 0.013$). Perceived value also shows a negative association with loyalty ($r = -0.375$, $p = 0.065$), approaching conventional significance thresholds. Trust exhibits a negative but non-significant correlation with loyalty ($r = -0.190$, $p = 0.362$), while AI awareness is also negatively related to loyalty without statistical significance ($r = -0.217$, $p = 0.298$).

These findings indicate that, within this pilot sample, higher evaluations of AI-related service attributes do not correspond to higher loyalty scores.

E. Exploratory Regression Analysis

An exploratory multiple regression model was estimated with AI awareness, perceived value, service quality, and trust as predictors of customer loyalty. The overall model is statistically significant ($F = 4.224$, $p = 0.012$), explaining 45.8% of the variance in loyalty ($R^2 = 0.458$). However, the direction of relationships contrasts with dominant theoretical expectations.

Both perceived value ($\beta = -0.465$, $p = 0.013$) and service quality ($\beta = -0.546$, $p = 0.006$) show significant negative associations with loyalty. AI awareness ($\beta = -0.062$, $p = 0.729$) and trust ($\beta = -0.016$, $p = 0.927$) are not significant predictors. Given the pilot nature of the dataset and measurement limitations, these results are interpreted cautiously as indicative of unexpected relational patterns rather than confirmatory evidence.

F. Group Comparisons and Visual Patterns

Independent-sample t-tests indicate no statistically significant gender differences in loyalty or trust. Descriptive comparisons across age groups suggest variability in loyalty, with higher mean loyalty among respondents aged 46–55 and lower mean loyalty among those aged 55+, although small cell sizes limit interpretability. Visual inspection of scatterplots and boxplots further reinforces the observed negative association between perceived value and loyalty and highlights demographic dispersion in loyalty scores.

V. DISCUSSION

This exploratory pilot study examined whether the commonly assumed positive link between AI-enabled retail experiences and customer loyalty is reflected in consumer perceptions in the UAE context. Overall, the results suggest that this assumption does not necessarily hold. In particular, the negative associations observed between perceived service quality, perceived value, and loyalty contrast with the dominant narrative that AI-driven efficiency and personalization naturally strengthen customer relationships.

One plausible interpretation is that consumers can evaluate AI-enabled retail features positively at a functional level while still remaining weakly attached to the retailer. In other words, higher convenience, smoother service, or more relevant offers may improve the transaction without strengthening relational loyalty. This distinction between functional satisfaction and emotional attachment is consistent with arguments in prior

research that AI performs well in analytical and mechanical tasks but may struggle to support the relational and emotional foundations associated with long-term loyalty, especially when human interaction is reduced or perceived as replaceable. In such settings, consumers may acknowledge service improvement but feel less connected to the brand itself.

Trust also appears central to interpreting the findings. Trust-related measures were relatively low in the pilot sample, and trust did not emerge as a positive predictor of loyalty. This pattern suggests that AI-enabled services may be experienced within a broader context of skepticism toward data use and automated decision-making. Even when AI-based features work effectively, concerns about privacy, transparency, and control may inhibit the translation of service convenience into loyalty. Existing literature often treats trust as a necessary condition for AI-driven personalization to generate relational outcomes; without that foundation, efficiency gains may remain transactional rather than loyalty-building.

The UAE retail environment provides a relevant lens for understanding these dynamics. The market combines rapid digitalization and strong retail competition with a culturally diverse consumer base and increasing attention to data protection and ethical technology use. In such a context, AI implementations that are perceived as overly automated, opaque, or intrusive may generate ambivalence or resistance rather than loyalty. The pilot findings therefore support the broader view that AI strategies focused primarily on hyper-personalization or operational efficiency can create relational risk if they are not aligned with customer expectations regarding transparency, choice, and human engagement.

At the same time, these findings must be interpreted carefully. The study is exploratory, based on a very small non-probability pilot sample, and reliability estimates for some composites were unstable. These conditions place clear limits on interpretation and do not support confirmatory claims. In this sense, the study should be read primarily as a pilot and instrument-diagnostic exercise intended to identify early patterns, possible tensions, and areas requiring further measurement refinement. Although negative directional relationships were observed across the exploratory analyses, these should be treated as preliminary signals requiring verification in larger and more stable samples.

Overall, the study contributes early context-specific insight into how AI-enabled retail features may be perceived in relation to loyalty in the UAE, while also highlighting the methodological challenges of examining these relationships in a pilot setting. Rather than concluding that AI-driven personalization or service quality reduces loyalty, the present findings more cautiously suggest that functional AI-enabled improvements may not automatically translate into stronger relational loyalty under conditions of low trust, privacy concern, or limited human engagement. Practically, this points to the importance of hybrid service models, transparent data practices, and explicit trust-building mechanisms alongside AI deployment. From a research perspective, the study motivates larger-scale investigations with refined measures to examine these relationships more robustly.

VI. LIMITATIONS AND FUTURE RESEARCH

This study is subject to several important limitations that should be considered when interpreting the findings. First, the analysis is based on a very small pilot sample ($n = 25$) collected using non-probability sampling. As a result, the findings are not intended to be generalizable to the broader UAE retail population. Rather, the sample is used to support an exploratory pilot objective focused on identifying early patterns, potential tensions, and measurement issues that can inform subsequent research design.

Second, the small sample size implies low statistical power, limited precision, and instability in estimated relationships. It also increases the risk that regression-based patterns may reflect sample-specific variation rather than robust underlying associations. For this reason, all inferential results in the study should be interpreted as tentative and descriptive rather than confirmatory.

Third, the cross-sectional design captures customer perceptions at a single point in time. Given the rapidly evolving nature of AI technologies and consumer expectations, longitudinal designs would be better suited to examine how trust, perceived value, and loyalty evolve as customers gain prolonged exposure to AI-enabled retail services. The present results should therefore be interpreted as indicative of contemporaneous associations rather than stable relationships.

Fourth, measurement instability was observed in composite constructs, particularly trust and loyalty. This raises concerns regarding construct consistency in the present pilot dataset and suggests that some items, including reverse-coded items, may require further refinement, testing, or re-specification in future work. The present study therefore serves partly as an instrument-testing stage, helping identify areas where construct definitions, item wording, and dimensional structure require improvement before larger-scale validation.

Future research should build on this pilot by employing larger and more diverse samples to improve robustness and generalizability. Future studies should also refine and validate the measurement instrument through item review, reverse-code verification, scale refinement, and construct validation procedures such as factor analysis where sample size permits. In addition, qualitative methods such as interviews or focus groups could help explain why AI-enabled convenience or personalization may not translate directly into loyalty at the relational level. More broadly, future research should examine trust, perceived intrusiveness, transparency, and human-AI service balance as central mechanisms shaping AI-loyalty relationships in culturally diverse and regulation-sensitive retail environments such as the UAE.

VII. CONCLUSION

This study offers an exploratory pilot examination of how AI-driven retail features relate to customer loyalty in the UAE. In this pilot sample, the findings suggest that higher perceived AI-enabled service quality and value did not clearly

correspond to stronger loyalty. Instead, the results point to a potential disconnect between functional efficiency delivered by AI and the relational foundations of customer loyalty, particularly in contexts where trust, transparency, and human-centric interaction matter.

While preliminary, these findings contribute to ongoing debates on the limits of AI-driven personalization and service automation in retail. The study highlights the importance of aligning AI deployment with trust-building practices and hybrid service models that preserve meaningful human engagement. Future research using larger samples and refined measures can further clarify the conditions under which AI supports, rather than undermines, sustainable customer loyalty.

REFERENCES

[1] Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2021). Customer experiences in smart retailing. *Technological Forecasting and Social Change, 173*, 121130. https://doi.org/10.1016/j.techfore.2021.121130

[2] Bibi, M., Aziz, W., Khan, I. R., & Nadeem, M. S. A. (2025). Twitter sentiment analysis using ensemble of sentic computing and a novel clustering optimisation algorithm based on combined proximity measures. *Journal of Information & Knowledge Management, 24*(6), 2550072. https://doi.org/10.1142/S0219649225500728

[3] Davenport, T. H., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

[4] Grewal, D., Roggeveen, A. L., & Nordfält, J. (2017). The future of retailing. *Journal of Retailing, 93*(1), 1–6. https://doi.org/10.1016/j.jretai.2016.12.008

[5] Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research, 21*(2), 155–172. https://doi.org/10.1177/1094670517752459

[6] Pillai, R., & Sivathanu, B. (2020). Adoption of AI-based chatbots for hospitality and tourism. *International Journal of Contemporary Hospitality Management, 32*(10), 3199–3226. https://doi.org/10.1108/IJCHM-04-2020-0259

[7] Prentice, C., Dominique Lopes, S., & Wang, X. (2020). The impact of artificial intelligence and employee service quality on customer engagement and loyalty. *Journal of Hospitality Marketing & Management, 29*(7), 803–822. https://doi.org/10.1080/19368623.2020.1722304

[8] Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence. *Journal of Marketing, 85*(1), 7–32. https://doi.org/10.1177/0022242920953847

[9] Rosado-Pulido, J., Muñoz-Leiva, F., & Liébana-Cabanillas, F. (2023). The paradox of personalization and privacy in AI-driven retailing. *Journal of Business Research, 156*, 113124. https://doi.org/10.1016/j.jbusres.2022.113124

[10] Shankar, V. (2018). How artificial intelligence (AI) is reshaping retailing. *Journal of Retailing, 94*(4), 1–5. https://doi.org/10.1016/j.jretai.2018.10.002

[11] Van der Aa, Z., Van Kollenburg, G., & Van Loenen, E. (2021). The role of trust in AI-driven customer experiences. *Journal of Consumer Behaviour, 20*(3), 345–358. https://doi.org/10.1002/cb.1874