Artificial Intelligence in Russia: Evaluating Yandex GPT-5 and GigaChat 2.0 MAX in Global Context

Arsenii Moshninov
Graduate School of Business
HSE University
Moscow, Russia
akmoshninov@edu.hse.ru

Mikhail Komarov
Graduate School of Business
HSE University
Moscow, Russia
mmkomarov@hse.ru

https://doi.org/10.63962/URBB4192

***Abstract*— Study evaluates Russian large language models Yandex GPT-5 and GigaChat 2.0 MAX against global leaders—ChatGPT 4o, Grok 3, and DeepSeek V3—in business-relevant tasks. This study uses a mixed-method approach, combining custom-designed tasks, LM Arena data, and standardized benchmarks to evaluate logic, mathematics, programming, content creation, and image generation. Russian models match global peers in logic and mathematics, lag slightly in programming and creativity due to resource constraints, and excel in native language tasks, partially outperforming Western models. Their performance aligns closely with international standards, with potential to lead in Russian tasks, reflecting Russia's strategic AI.**

Keywords—AI development, Russian AI, artificial intelligence, Yandex GPT-5, GigaChat 2.0 MAX

I. Introduction

The global artificial intelligence (AI) landscape has entered an era of geopolitical stratification, wherein national AI systems increasingly embody regional technological priorities and cultural contexts. Russia's pursuit of technological sovereignty has caused the rapid advancement of domestic large language models (LLMs), notably Yandex GPT-5 and GigaChat 2.0 MAX, developed by Russia's leading IT companies, Yandex and Sber, respectively. Yandex GPT- 5, created by the search engine giant Yandex, is optimized for Russian language processing and integrates with the Alice voice assistant [1]. GigaChat 2.0 MAX, developed by the fintech leader Sber, excels in Russian language contexts [2]. These advancements occur amidst international sanctions and an increased emphasis on import substitution. They foster a unique AI innovative ecosystem that balances limited resources with strategic investments in technology and linguistics [3,4,5].

Recent analyses by CNA Russia Studies indicate that 78% of Russian AI research funding is channeled through state-aligned entities, emphasizing applications that ensure information control and linguistic autonomy [3]. This state-driven approach contrasts sharply with the commercial paradigms prevalent in the West, yet it delivers significant outcomes. For instance, GigaChat 2.0 MAX achieves an 80% accuracy rate on the Massive Multitask Language Understanding (MMLU) benchmark within Russian contexts, surpassing certain Chinese models in regional assessments. Similarly, Yandex's integration of the Alice voice assistant with GPT-5 Pro highlights Russia's tailored approach to human-computer interaction, with a focus on optimizing Cyrillic script processing and Slavic linguistic structures [6,7].

Western advancements show scale and versatility, while Russian LLMs like Yandex GPT-5 prioritize efficiency, achieving 92% of GPT-4's coding accuracy with 40% fewer computational resources—a critical advantage given Russia's reliance on sanctioned hardware alternatives [8,9,10].

Russia's strategic AI positioning blends technological resilience with cultural specificity. The 2024 CSET analysis shows only 12% of Russian studies address cross-cultural evaluation, compared to 34% in U.S. research [4,5].

Comprehensive comparisons between Russian and global LLMs in business applications are scarce. Existing research on Russian AI, like Yandex GPT-5 and GigaChat 2.0 MAX, is outdated due to rapid innovation, lacking rigorous scrutiny. This study addresses this gap by systematically comparing their strengths and limitations for organizational performance

This research contributes to AI in business management by introducing a novel evaluation framework accounting for cultural and linguistic variations, comparing Russian and Western LLMs to describe global AI dynamics, and providing actionable insights for enterprises and policymakers in cross-cultural contexts [11,12,13].

II. Methodology

This study adopts a mixed-methods framework to assess the performance of five LLMs in business-relevant tasks—ChatGPT 4o by OpenAI [14], Grok 3 by xAI, a company associated with Elon Musk [15], DeepSeek V3, a non-commercial model from a Chinese startup [16], GigaChat 2.0

Artificial intelligence was applied to facilitate the formatting and stylistic aspects of this short paper.

MAX by Sber, a fintech leader excelling in Russian-language contexts [2], and Yandex GPT-5 by Yandex, a search engine giant optimized for Russian language processing [1]. The approach combines custom-designed tasks, LM Arena data, and standardized benchmarks for a robust evaluation, as illustrated in the methodology flowchart (Figure 1) [17].

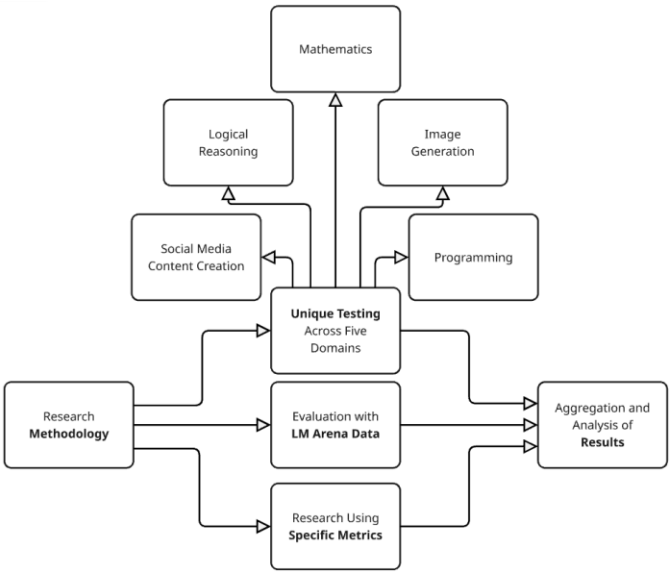


Fig. 1. Flowchart with research methodology

We designed tasks across five categories: logical reasoning (e.g., letter counting, scored 0 or 1) [18], mathematics (e.g., equation solving, partial credit), programming (JavaScript 'Snake' game, 0–2 for functionality), social media (Instagram post, 1–5 for creativity), and image generation (futuristic city, assessed by three for accuracy, averaged).

The authors acknowledge uncertainty regarding whether the results may have been affected by the service's concurrent load, potentially reducing response accuracy and task-solving efficacy, thus constituting a study limitation.

LM Arena comparison utilizes the platform's Elo rating system to rank ChatGPT 4o, Grok 3, and DeepSeek V3 based on user interactions across various Russian, English tasks. For models not included in LM Arena—GigaChat 2 MAX and YandexGPT 5—simulated pairwise comparisons or analogous methods will estimate their performance relative to listed models [17].

Standardized benchmarks provide objective metrics: MMLU (Russian and English) for language comprehension, GSM8K and MATH for mathematical reasoning, HumanEval for coding, and IFEVAL (Russian and English) for instruction-following [17,19,20,21,22].

Evaluation and comparison integrate results from all components. Custom task scores (binary or scaled), LM Arena rankings, and benchmark outcomes are aggregated to compare model performance holistically, identifying strengths and limitations for business uses like customer service and content creation.

III. Results

All models demonstrated proficiency in logical reasoning and mathematical tasks, achieving maximum scores across the board (e.g., 5/5 points each for ChatGPT 4o, Grok 3, DeepSeekV3, Yandex GPT, and GigaChat 2 MAX in Table I), which indicates a strong baseline capability in structured problem-solving based on accuracy and efficiency. In programming tasks, ChatGPT 4o, Grok 3, and DeepSeekV3 distinguished themselves with fully functional solutions that met all specified requirements, showcasing robust code generation abilities and earning top marks (e.g., 4/4 points each) for executable and efficient code. Conversely, Yandex GPT's website design appeared outdated, lacking modern aesthetic and functional standards, scoring lower (e.g., 2/4 points) due to its less effective output, while GigaChat 2 MAX failed to produce operational game code, yielding non-executable results and receiving minimal points (e.g., 1/4). For social media content creation, DeepSeekV3 exhibited exceptional creativity, crafting a post that excelled in engagement, relevance, and narrative depth, outperforming its peers in this qualitative assessment and securing the highest score (e.g., 5/5 points), while others scored based on comparative quality (e.g., 3–4/5). In the image generation task, ChatGPT 4o accurately integrated text and imagery as per the prompt, earning a high score (e.g., 5/5 points) for precision and coherence. Yandex GPT and GigaChat 2 MAX achieved moderate success, though their results were marred by inaccuracies in text rendering and visual coherence, resulting in moderate scores (e.g., 3–4/5 points). Notably, DeepSeekV3 lacks image generation functionality entirely, a limitation inherent to its current design, rendering it unable to compete in this domain and thus scoring 0/5. Despite these variations, the performance gap between Russian models (GigaChat 2 MAX and Yandex GPT) and global models was not substantial, suggesting comparable overall competence as reflected in Table I's aggregate scores. However, ChatGPT 4o and Grok 3 slightly led the evaluation, due to their consistent performance, balancing technical precision with creative adaptability, achieving higher total scores (e.g., 18–19/20 vs. 14–16/20). Detailed results and scores for each model across the evaluated tasks are presented in Appendix A, a brief version in Table I.

TABLE I. BRIEF VERSION AI TEST

Task Type	ChatGPT-4o	Grok-3	DeepSeekV3	GigaChat 2 MAX	Yandex GPT
Logical reasoning tasks	5	5	5	5	5
Mathematical tasks	3	3	3	3	3
Programming	4	4	4	2	2
SMM	4	4	4	5	3
Image generation	5	4	3	4	4
Total score	**21**	**20**	**17**	**16**	**17**

Fig. 2. Brief version AI test, Table I

Chatbot Arena data (no Russian LLMs) shows ChatGPT-4o ranking 2nd overall (1466), Grok-3 4th (1404), and DeepSeek-V3 5th (1370). In Russian tasks, ChatGPT-4o and Grok-3 tie for 1st (1430, 1426), DeepSeek-V3 ranks 3rd (1376). Grok-3 excels in creativity (2nd, 1406), ChatGPT-4o in coding (1st, 1425), while DeepSeek-V3 lags in math (6th, 1341), indicating ChatGPT-4o's versatility across tasks [17] (Appendix B).

Benchmark analysis reveals GigaChat 2 MAX's competitive edge in Russian-language tasks, scoring 80.46 on MMLU (RU), surpassing GPT-4o (80.00) and DeepSeek-V3 (73.74). In English tasks, GPT-4o leads with 88.70 on MMLU (EN), followed by GigaChat 2 MAX (86.00) and DeepSeek-V3 (85.24). GigaChat 2 MAX excels in instruction-following for Russian (IFEVAL RU: 83.62), while DeepSeek-V3 dominates in coding (HumanEval: 91.46) and English instruction-following (IFEVAL EN: 92.21). These results highlight the strength of Russian LLMs in native language contexts, positioning GigaChat 2 MAX as a formidable contender globally [23] (Appendix C).

IV. Conclusion

This study has demonstrated that the latest Russian LLMs, Yandex GPT-5 and GigaChat 2.0 MAX, exhibit performance levels closely aligned with global models such as DeepSeek V3, ChatGPT 4o, and Grok 3. The research reveals that Russian LLMs achieve competitive outcomes across diverse business-relevant tasks. Notably, their proficiency in logical reasoning and mathematics matches that of global counterparts, though they lag slightly in programming and creative domains, reflecting resource constraints rather than inherent limitations.

In Russian-language tasks, global models showcased varying strengths based on LM Arena data (excluding Russian models). ChatGPT 4o and Grok 3 tied for the highest performance, achieving Elo ratings of 1430 and 1426, respectively, while DeepSeek V3 trailed at 1376. This indicates that ChatGPT 4o and Grok 3 excel in versatility and adaptability across linguistic contexts, positioning them as leaders among global models for Russian tasks [13].

Comparatively, Russian models partially outperform Western counterparts in native language benchmarks. GigaChat 2.0 MAX, for instance, scored 80.46 on MMLU (RU), surpassing ChatGPT 4o (80.00) and DeepSeek V3 (73.74), and excelled in instruction-following (IFEVAL RU: 83.62). These results underscore the tailored efficacy of Russian models in their linguistic domain, driven by a focus on cultural and functional specificity.

The unique geopolitical and geotechnical motivations of Russian developers—emphasizing technological sovereignty and linguistic autonomy—suggest a trajectory where Russian models may soon outpace global analogs in Russian-language tasks. With state-driven investments and progress evident in current benchmarks, models like GigaChat 2.0 MAX are steadily advancing toward this potential, leveraging regional priorities to bridge performance gaps.

Globally, Russian LLMs stand roughly equivalent to their international peers, mirroring the developmental pace of Western AI systems. This parity highlights Russia's strategic resilience in AI innovation despite sanctions and limited resources. This study does not address the data infrastructure challenges of LLMs, which remain beyond its scope. Future research should explore longitudinal comparisons and the broader impact of national strategies on AI ecosystems, offering deeper insights into their implications for global business management.

References

[1] Yandex. (n.d.). *Alice voice assistant*. Retrieved April 22, 2025, from https://alice.yandex.ru/

[2] Sber. (n.d.). *GigaChat*. Retrieved April 22, 2025, from https://giga.chat/

[3] Center for Naval Analyses. (2024). *Artificial intelligence and autonomy in Russia*. Retrieved April 4, 2025, from https://www.cna.org/centers-and-divisions/cna/sppp/russia-studies/artificial-intelligence-and-autonomy-in-russia

[4] Geneva Internet Platform. (2024, February 11). *Russia struggles to catch up in global AI race*. Retrieved April 4, 2025, from https://dig.watch/updates/russia-struggles-to-catch-up-in-global-ai-race

[5] Popkova, E. G., & Stefanovic, M. (2024). [TRENDS OF THE AI ECONOMY IN RUSSIA]. Journal of Artificial Intelligence, 1(1), 1–10. https://jai.aspur.rs/archive/v1/n1/1.pdf

[6] Evrim Ağacı. (2025, February 25). *YandexGPT-5 Pro revolutionizes AI with Alice integration*. Retrieved April 4, 2025, from https://evrimagaci.org/tpg/yandexgpt-5-pro-revolutionizes-ai-with-alice-integration-227608

[7] The Hans India. (2025, March 13). *Sber presents new neural network GigaChat 2.0*. Retrieved April 4, 2025, from https://www.thehansindia.com/business/sber-presents-new-neural-network-gigachat-20-953634

[8] Konaev, M., & Gilli, A. (2020). *Russian AI research 2010 to 2018*. Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/CSET-Russian-AI-Research-2010-to-2018-2.pdf

[9] Future Skills Academy. (2025, February 25). *GPT-5 vs GPT-4*. Retrieved April 4, 2025, from https://futureskillsacademy.com/blog/gpt-5-vs-gpt-4/

[10] Fello AI. (2024, August). *Claude AI: Everything you need to know*. Retrieved April 4, 2025, from https://felloai.com/2024/08/claude-ai-everything-you-need-to-know/

[11] Chen, D., Esperança, J. P., & Wang, S. (2022). The impact of artificial intelligence on firm performance: An application of the resource-based view to e-commerce firms. *Frontiers in Psychology, 13*, Article 884830. https://doi.org/10.3389/fpsyg.2022.884830

[12] DataCube Research. (2024, June). *Russia generative AI market: Analysis 2019-2032* (Report AI4212). Niche Industry Monitor.

[13] Business Insider. (2024). *US, China compete for AI dominance while Russia's model lags behind*. Retrieved April 4, 2025, from https://www.businessinsider.com/us-china-compete-ai-dominance-while-russia-model-lags-behind-2025-2

[14] OpenAI. (n.d.). *ChatGPT*. Retrieved April 22, 2025, from https://chatgpt.com/

[15] xAI. (n.d.). *Grok*. Retrieved April 22, 2025, from https://grok.com/

[16] DeepSeek. (n.d.). *DeepSeek chat*. Retrieved April 22, 2025, from https://chat.deepseek.com

[17] Chatbot Arena. (2024). *LM Arena Leaderboard*. Retrieved April 4, 2025, from https://arena.lmsys.org/

[18] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv*. https://arxiv.org/abs/2009.03300

[19] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv*. https://arxiv.org/abs/2110.14168

[20] Jagtap, A. D., Shin, Y., Kawaguchi, K., & Karniadakis, G. E. (2022). Deep Kronecker neural networks: A general framework for neural networks with adaptive activation functions. *Neurocomputing, 468*, 165–180.

[21] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B.,

Gray, S., … Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv*. https://arxiv.org/abs/2107.03374

[22] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[23] T-J.ru. (2025, March 13). *Sber GigaChat 2*. Retrieved April 4, 2025, from https://t-j.ru/news/sber-gigachat-2/

APPENDIX A

Type	Task	ChatGPT 4o	Grok 3	DeepSeekV3	GigaChat 2 MAX	Yandex GPT
Logical reasoning tasks	Letter count in word	1	1	1	1	1
	Next number in sequence	1	1	1	1	1
	Logical deduction about tails	1	1	1	1	1
	Calculate average speed	1	1	1	1	1
	Anagram verification	1	1	1	1	1
Total Logical		**5**	**5**	**5**	**5**	**5**
Mathematical tasks	Solve linear equation	1	1	1	1	1
	Find function derivative	1	1	1	1	1
	Prove irrationality	1	1	1	1	1
Total mathematical		**3**	**3**	**3**	**3**	**3**
Programming	Create course website HTML	2	2	2	2	1
	Develop JavaScript snake game	2	2	2	0	1
Total programming		**4**	**4**	**4**	**2**	**2**
SMM	Write eco-product Instagram post	**4**	**4**	**5**	**3**	**3**
Image generation	Generate futuristic city image	**5**	**4**	**0**	**3**	**4**
Total score		**21**	**20**	**17**	**16**	**17**

Fig. 3. Detailed results and scores for each model across the evaluated tasks

APPENDIX B

Category: Russian

Apply filter: Style Control, Show Deprecated

Russian Prompts

#models: 193 (87%) #votes: 298,922 (11%)

Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	0	Gemini-2.5-Pro-Exp-03-25	1467	+27/-28	505	Google	Proprietary
1	1	GPT-4.5-Preview	1430	+23/-18	1323	OpenAI	Proprietary
1	1	ChatGPT-4o-latest (2025-03-26)	1430	+36/-30	394	OpenAI	Proprietary
1	1	Grok-3-Preview-02-24	1426	+21/-20	1235	xAI	Proprietary
2	3	Gemini-2.0-Pro-Exp-02-05	1416	+14/-13	2402	Google	Proprietary
2	3	Gemini-2.0-Flash-Thinking-Exp-01-21	1404	+12/-14	2724	Google	Proprietary
3	2	DeepSeek-V3-0324	1376	+32/-35	318	DeepSeek	MIT

Fig. 4. LM Arena data – Russian Tasks score [17].

Category: Math

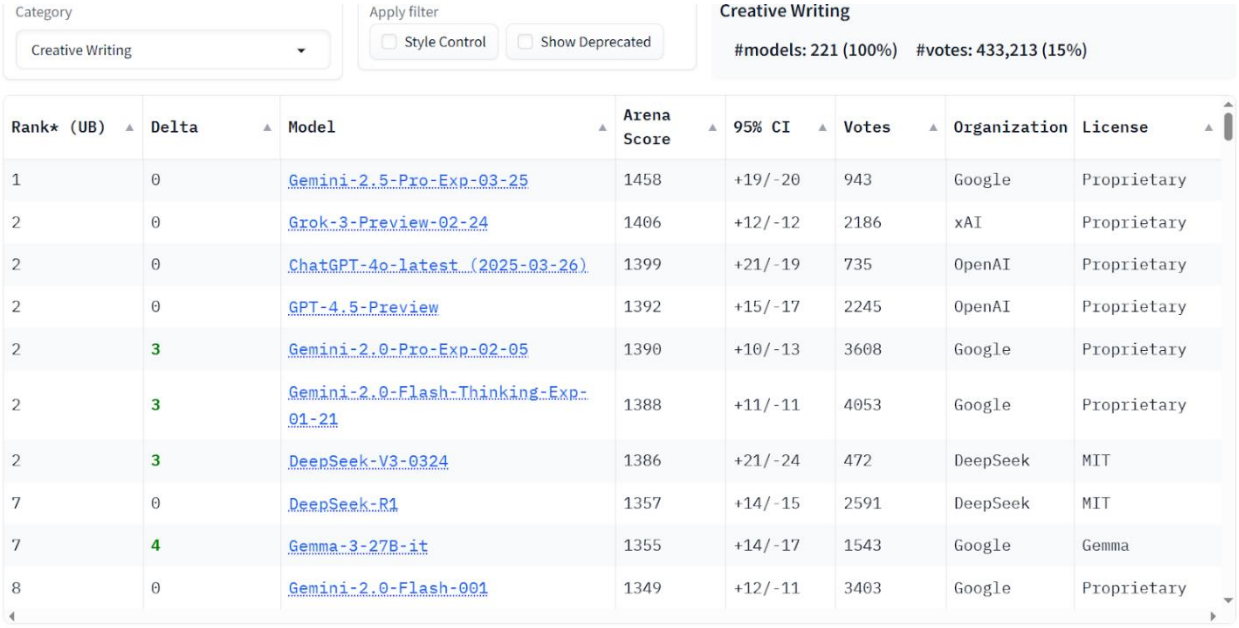
Apply filter: Style Control, Show Deprecated

Math

#models: 219 (99%) #votes: 378,834 (13%)

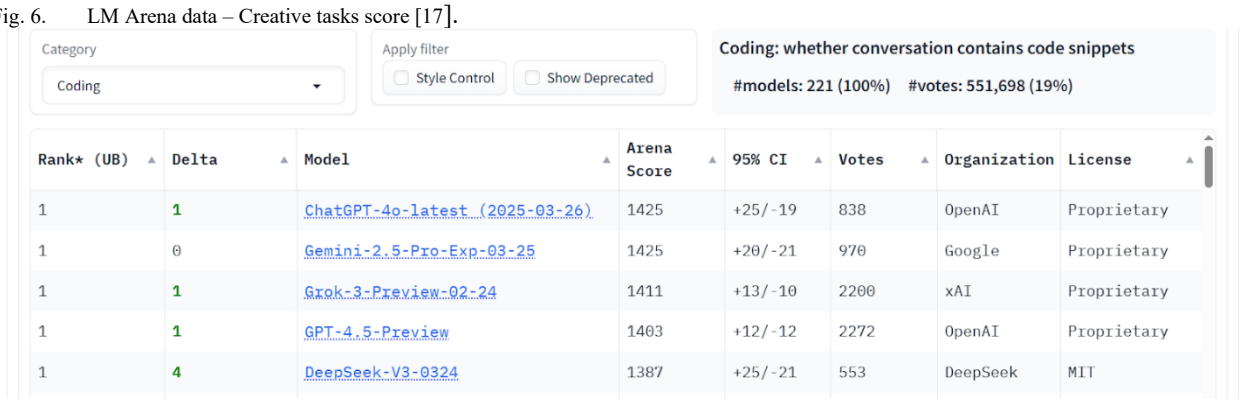
Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	0	Gemini-2.5-Pro-Exp-03-25	1414	+22/-21	606	Google	Proprietary
2	0	GPT-4.5-Preview	1377	+15/-12	1248	OpenAI	Proprietary
2	5	DeepSeek-R1	1359	+17/-15	1339	DeepSeek	MIT
2	6	o1-2024-12-17	1355	+14/-12	2960	OpenAI	Proprietary
2	12	o3-mini-high	1354	+15/-15	1678	OpenAI	Proprietary
2	3	DeepSeek-V3-0324	1341	+40/-29	284	DeepSeek	MIT
3	14	o3-mini	1348	+10/-12	2574	OpenAI	Proprietary
3	-1	Grok-3-Preview-02-24	1348	+13/-15	1214	xAI	Proprietary
3	2	Gemini-2.0-Flash-Thinking-Exp-01-21	1341	+13/-11	2734	Google	Proprietary
3	8	o1-preview	1339	+9/-10	5052	OpenAI	Proprietary
3	2	Gemini-2.0-Pro-Exp-02-05	1332	+12/-11	2381	Google	Proprietary
3	-1	ChatGPT-4o-latest (2025-03-26)	1330	+24/-26	516	OpenAI	Proprietary

Fig. 5. LM Arena data – Mathematical tasks score [17].



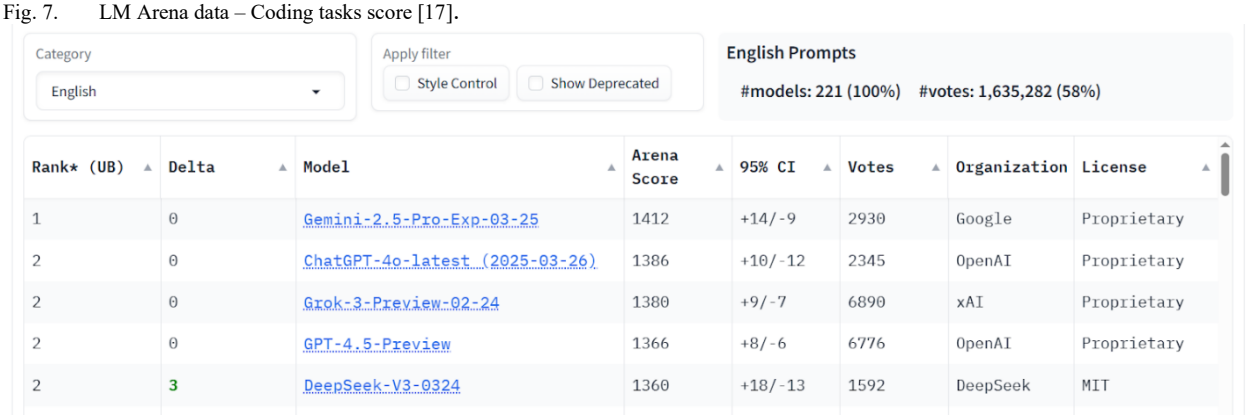
Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	0	Gemini-2.5-Pro-Exp-03-25	1458	+19/-20	943	Google	Proprietary
2	0	Grok-3-Preview-02-24	1406	+12/-12	2186	xAI	Proprietary
2	0	ChatGPT-4o-latest (2025-03-26)	1399	+21/-19	735	OpenAI	Proprietary
2	0	GPT-4.5-Preview	1392	+15/-17	2245	OpenAI	Proprietary
2	3	Gemini-2.0-Pro-Exp-02-05	1390	+10/-13	3608	Google	Proprietary
2	3	Gemini-2.0-Flash-Thinking-Exp-01-21	1388	+11/-11	4053	Google	Proprietary
2	3	DeepSeek-V3-0324	1386	+21/-24	472	DeepSeek	MIT
7	0	DeepSeek-R1	1357	+14/-15	2591	DeepSeek	MIT
7	4	Gemma-3-27B-it	1355	+14/-17	1543	Google	Gemma
8	0	Gemini-2.0-Flash-001	1349	+12/-11	3403	Google	Proprietary

Fig. 6. LM Arena data – Creative tasks score [17].



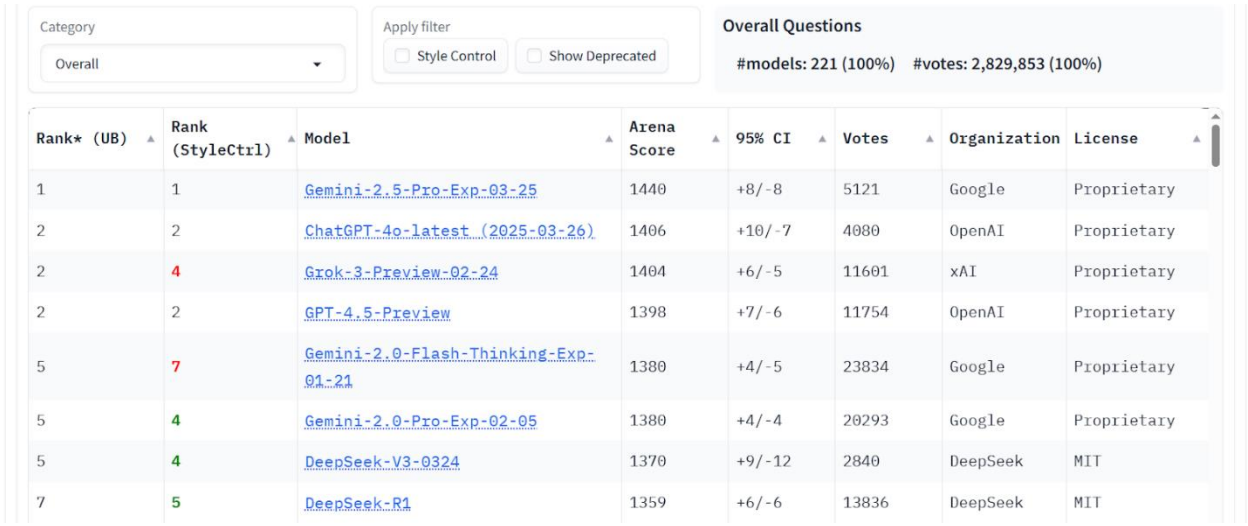
Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	1	ChatGPT-4o-latest (2025-03-26)	1425	+25/-19	838	OpenAI	Proprietary
1	0	Gemini-2.5-Pro-Exp-03-25	1425	+20/-21	970	Google	Proprietary
1	1	Grok-3-Preview-02-24	1411	+13/-10	2200	xAI	Proprietary
1	1	GPT-4.5-Preview	1403	+12/-12	2272	OpenAI	Proprietary
1	4	DeepSeek-V3-0324	1387	+25/-21	553	DeepSeek	MIT

Fig. 7. LM Arena data – Coding tasks score [17].



Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	0	Gemini-2.5-Pro-Exp-03-25	1412	+14/-9	2930	Google	Proprietary
2	0	ChatGPT-4o-latest (2025-03-26)	1386	+10/-12	2345	OpenAI	Proprietary
2	0	Grok-3-Preview-02-24	1380	+9/-7	6890	xAI	Proprietary
2	0	GPT-4.5-Preview	1366	+8/-6	6776	OpenAI	Proprietary
2	3	DeepSeek-V3-0324	1360	+18/-13	1592	DeepSeek	MIT

Fig. 8. LM Arena data – English tasks score [17].



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1440	+8/-8	5121	Google	Proprietary
2	2	ChatGPT-4o-latest (2025-03-26)	1406	+10/-7	4080	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1404	+6/-5	11601	xAI	Proprietary
2	2	GPT-4.5-Preview	1398	+7/-6	11754	OpenAI	Proprietary
5	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+4/-5	23834	Google	Proprietary
5	4	Gemini-2.0-Pro-Exp-02-05	1380	+4/-4	20293	Google	Proprietary
5	4	DeepSeek-V3-0324	1370	+9/-12	2840	DeepSeek	MIT
7	5	DeepSeek-R1	1359	+6/-6	13836	DeepSeek	MIT

Fig. 9. LM Arena data – Overall score [17].

APPENDIX C

Category	Benchmark Name	GigaChat 2 MAX	Qwen 2.5 72B	Llama 3.3 70B	GPT-4o	DeepSeek-V3
General Knowledge	MMLU (RU)	80,46	78,30	65,08	80,00	73,74
	MMLU (EN)	86,00	83,85	78,57	88,70	85,24
Mathematics	GSM8K	95,68	95,07	92,87	95,00	94,99
	MATH	77,26	78,74	62,80	76,60	85,48
Coding	HumanEval	87,20	86,60	86,00	84,00	91,46
Instruction Following	IFEVAL (RU)	83,62	84,27	75,12	80,24	84,37
	IFEVAL (EN)	89,99	90,43	90,83	88,51	92,21

Fig. 10. tests of models based on metrics by task category [23].