Cross-Lingual Mathematical Reasoning in LLMs: Evaluating Performance on Icelandic vs. English Problems

Hafsteinn Einarsson
Department of Computer Science, University of Iceland
Reykjavik, Iceland
hafsteinne@hi.is

Abstract

We investigate whether large language models (LLMs) exhibit performance differences when solving mathematical problems presented in a low-resource language (Icelandic) versus a high-resource language (English). Using 847 multiple-choice problems from the Icelandic Mathematics Competition corpus (STAK), we evaluate two state-of-the-art models (Gemini-3-Flash-Preview and GPT-5.4-mini) in both multiple-choice (MC) and open-ended (OE) formats, with correctness determined by a three-judge quorum (Gemini-3-Flash, GPT-5.4-mini, Claude Sonnet 4.6) achieving 97.6% unanimous agreement. Our results reveal significant cross-lingual performance gaps that vary by model: Gemini-3-Flash shows a consistent English advantage of 2.4–10.0 percentage points across both evaluation modes, while GPT-5.4-mini exhibits no significant language effects. Notably, GPT-5.4-mini demonstrates a substantial MC deficit, achieving only 42% in that format despite reaching 69–71% accuracy on OE problems. Analysis of answer patterns reveals a strong option position bias in GPT-5.4-mini, with systematic over-selection of option B and under-selection of option D. These findings suggest that language does affect LLM mathematical reasoning for some models, but the effect is model-dependent and interacts with evaluation format, with implications for deploying LLMs in educational contexts for speakers of low-resource languages.

Keywords: mathematical reasoning, cross-lingual evaluation, low-resource languages, Icelandic, large language models, LLM evaluation

1. Introduction

LLM training corpora are dominated by English. English accounts for approximately 92.65% of GPT-3's training tokens (OpenAI, 2023) and 89.70% of LLaMA2's pre-training data (Touvron et al., 2023). Recent efforts such as OLMo 3 apply language filters that retain only English text (Olmo et al., 2025). Low-resource languages receive minimal representation, raising questions about whether LLMs can reason equally well in those languages.

Mathematical reasoning provides a controlled test of cross-lingual ability. The underlying concepts are language-independent; only the linguistic framing changes between translations. If an LLM has learned to reason mathematically, its performance should hold across languages.

Icelandic, with approximately 370,000 speakers, sits firmly in the low-resource category. As a North Germanic language in the Indo-European family, Icelandic shares deep roots with English through their common Germanic ancestry but has preserved a significantly more complex morphological system: four grammatical cases, three genders, and extensive noun and verb inflection. Unlike other Nordic languages (Norwegian, Danish, Swedish), Icelandic has resisted English lexical borrowing, maintaining a largely native vocabulary. These characteristics make Icelandic an informative test case: it shares enough structural overlap with English that mathematical terminology is broadly translatable, yet its morphological complexity and lower digital representation create meaningful processing challenges for LLMs. Results may therefore be indicative of performance on other morphologically rich, low-resource European languages.

The resulting training-data disparity could reduce comprehension of Icelandic problem statements or degrade mathematical reasoning on Icelandic text. This raises a direct question: *does the language of mathematical problem presentation affect LLM performance?*

We evaluate two models on the STAK dataset (Einarsson et al., 2026), which comprises 847 problems from Icelandic mathematics competitions (1984–2025) covering algebra, geometry, number theory, and combinatorics. Each problem was machine-translated to English, and a random sample of 100 translations was manually verified by a native Icelandic speaker with professional English proficiency, enabling direct cross-lingual comparison in both multiple-choice and open-ended formats.

2. Related Work

LLM Mathematical Reasoning Evaluation Recent benchmarks have enabled systematic evaluation of LLM mathematical reasoning capabilities. GSM8K (Cobbe et al., 2021) introduced 8,500 grade-school math problems requiring two to eight reasoning steps, establishing verification-based training as a strategy for improving mathematical accuracy. The MATH benchmark (Hendrycks et al.,

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 89–95 May 11, 2026.

2021) raised the bar with 12,500 competition-level problems where models initially achieved just 6.9% accuracy. For multilingual evaluation, MGSM (Shi et al., 2023) translated 250 GSM8K problems into ten typologically diverse languages, demonstrating that chain-of-thought reasoning transfers across languages with increasing model scale. However, MGSM translates from English to other languages, whereas we translate from a low-resource source (Icelandic) to English. These benchmarks primarily target English, leaving cross-lingual mathematical reasoning understudied.

Cross-Lingual NLP Performance Gaps Prior work has documented performance disparities between high-resource and low-resource languages across various NLP tasks. Chen et al. (2024) found that open-source LLMs suffer significant degradation on multilingual mathematical reasoning, particularly for low-resource languages, and proposed training on parallel multilingual corpora to close these gaps. MMLU-ProX (Xuan et al., 2025), covering 29 languages with 11,829 parallel questions, reveals large performance gaps between high-resource and low-resource languages. The "Mother Tongue Effect" (Fabbri et al., 2025) captures this phenomenon: models perform differently on culturally grounded reasoning depending on whether problems are presented in native languages or English translations. Mathematical reasoning differs: the underlying task is ostensibly language-independent so one might expect to not see a mother tongue effect.

Low-Resource Language Evaluation The evaluation of LLMs on low-resource languages faces challenges including limited benchmark availability and potential contamination of translated test sets. Data contamination might inflate benchmark performance, masking true generalization capabilities (Deng et al., 2024). Our STAK dataset has not been publicly released before this work, reducing contamination risk and providing a cleaner signal of genuine mathematical reasoning ability. Regarding translation reliability, Thellmann et al. (2024) found that machine-translated benchmarks can serve as reliable proxies for human evaluation, particularly when translating into well-resourced languages like English.

3. Methodology

3.1. Dataset

Our evaluation dataset consists of 847 multiple-choice problems, with each problem available in both Icelandic (original) and English (machine-translated). The problems are sourced from the STAK collection (Einarsson et al., 2026) and span competition years 1984–2025. Difficulty levels range from 1 to 10, with categories including algebra, geometry, number theory, and combinatorics. Each problem has four answer choices. The English translations were generated using Gemini-3-Flash and a random sample of 100 translations was verified by a native Icelandic speaker to confirm semantic faithfulness to the original problems.

3.2. Models Under Evaluation

We evaluated two state-of-the-art LLMs:

1. **Gemini-3-Flash-Preview** (Google), a high-performance model optimized for speed and accuracy
2. **GPT-5.4-mini** (OpenAI), a recent generation of OpenAI's efficient reasoning model

Both models were evaluated using identical prompts and evaluation code. To ensure model-agnostic evaluation, all API calls were routed through OpenRouter, eliminating any provider-specific differences in request handling.

3.3. Evaluation Modes

Each model was tested in two distinct evaluation modes:

- **Multiple-choice (MC)**: The model selects from provided answer options
- **Open-ended (OE)**: The model generates the answer independently

This dual-mode approach allows us to examine whether multiple-choice scaffolding mitigates or exacerbates language effects.

3.4. LLM Judge Quorum

To mitigate self-enhancement bias, where an LLM judge may favor outputs similar to its own (Gu et al., 2024), we employed a three-judge quorum comprising models from three independent providers: Gemini-3-Flash (Google), GPT-5.4-mini (OpenAI), and Claude Sonnet 4.6 (Anthropic). Each response was independently evaluated by all three judges, with the final correctness verdict determined by majority vote (2-of-3 agreement). The judges assessed whether each model response correctly solved the given problem, accounting for mathematical equivalence of numerical answers and correct selection of multiple-choice options. This approach builds on Zheng et al. (2023), who demonstrated that strong LLM judges achieve over 80% agreement with human preferences. The three judges achieved 97.6% unanimous agreement across all

Model	Condition	Acc.	95% CI
Gemini	IS, MC	78.51%	[75.8–81.2]
	EN, MC	88.55%	[86.3–90.7]
	IS, OE	87.49%	[85.2–89.6]
	EN, OE	89.85%	[87.7–91.9]
GPT-5.4-mini	IS, MC	42.50%	[39.1–45.8]
	EN, MC	41.79%	[38.5–45.1]
	IS, OE	68.83%	[65.8–71.9]
	EN, OE	70.96%	[67.9–73.9]

Table 1: Accuracy with 95% bootstrap confidence intervals, determined by three-judge quorum. $n = 847$ problems per condition.

6,776 evaluated responses, with 99.2% pairwise agreement between Claude and Gemini, 98.3% between Claude and GPT-5.4-mini, and 97.7% between Gemini and GPT-5.4-mini. This high inter-judge agreement provides strong evidence that evaluation results are robust and not driven by any single judge's biases.

3.5. Statistical Analysis

Since our outcomes are binary (correct/incorrect), we employed McNemar's test for paired binary comparisons to assess the significance of language effects within each model/mode combination, with the null hypothesis that the proportion of correct responses is equal for Icelandic and English presentations. For discordant pair counts below 25, we used the exact binomial test; otherwise, the chi-squared approximation. Bootstrap confidence intervals (95%, 10,000 resamples) were computed to estimate accuracy uncertainty.

4. Results

4.1. Overall Performance

Table 1 presents the accuracy for each model across all conditions with 95% bootstrap confidence intervals.

Figure 1 visualizes the overall performance across all conditions.

Gemini-3-Flash substantially outperforms GPT-5.4-mini across all conditions. Both models perform better in English than Icelandic. The OE mode shows higher accuracy than MC for both models, with the gap being dramatically larger for GPT-5.4-mini.

4.2. Statistical Significance of Language Effects

McNemar's tests comparing English versus Icelandic accuracy ($\alpha = 0.05$) reveal a clear split between models. For Gemini-3-Flash, the English advantage is significant in both MC mode ($p < 0.001$) and OE mode ($p = 0.002$). GPT-5.4-mini shows no significant language effect in either mode, with MC ($p = 0.703$) and OE ($p = 0.098$) both failing to reject the null hypothesis. One model exhibits robust sensitivity to presentation language across formats; the other appears indifferent to it.

4.3. Language Effect Analysis

Figure 2 visualizes the language effect (English accuracy minus Icelandic accuracy) for each model and evaluation mode.

For Gemini-3-Flash, the English advantage is +10.04 percentage points in MC mode and +2.36 percentage points in OE mode, both statistically significant. For GPT-5.4-mini, language effects are negligible: -0.71 percentage points in MC and +2.13 percentage points in OE, neither statistically significant.

5. Discussion

Our findings align with prior work documenting cross-lingual performance gaps (Chen et al., 2024; Xuan et al., 2025), though the gaps we observe for Icelandic are smaller than the maximum differences reported for other low-resource languages and earlier generation of LLMs.

5.1. Model Comparison

The performance gap between Gemini-3-Flash and GPT-5.4-mini is large. In MC mode, Gemini achieves 78.51% (IS) versus GPT-5.4-mini's 42.50% (IS), a gap of 36 percentage points. In OE mode, this gap narrows to approximately 19 percentage points (87.49% vs. 68.83%). The larger gap in MC mode suggests Gemini handles that format more effectively, while GPT-5.4-mini's reasoning capabilities manifest more clearly in open-ended settings.

5.2. Language Effect Patterns

The language effects are model-dependent. Gemini-3-Flash shows consistent, significant English advantages across both evaluation modes, suggesting systematic processing advantages for English mathematical text. GPT-5.4-mini shows no significant language effects in either mode, suggesting that for this model, other factors dominate over language effects.

5.3. Multiple-Choice Deficit and Option Position Bias

GPT-5.4-mini shows dramatic performance degradation in MC mode. The model achieves only

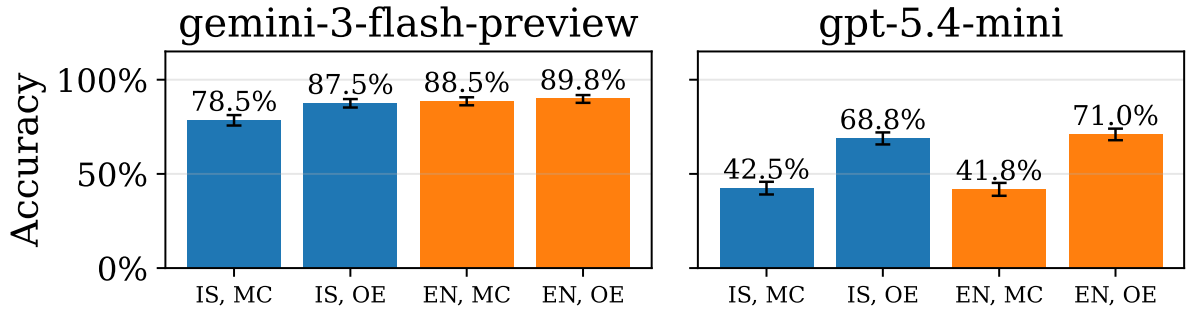

Figure 1: Overall accuracy by model, language, and evaluation mode with 95% bootstrap confidence intervals. Left: Gemini-3-Flash shows strong performance with a consistent English advantage. Right: GPT-5.4-mini exhibits substantially lower accuracy in MC mode but performs better in OE evaluation.

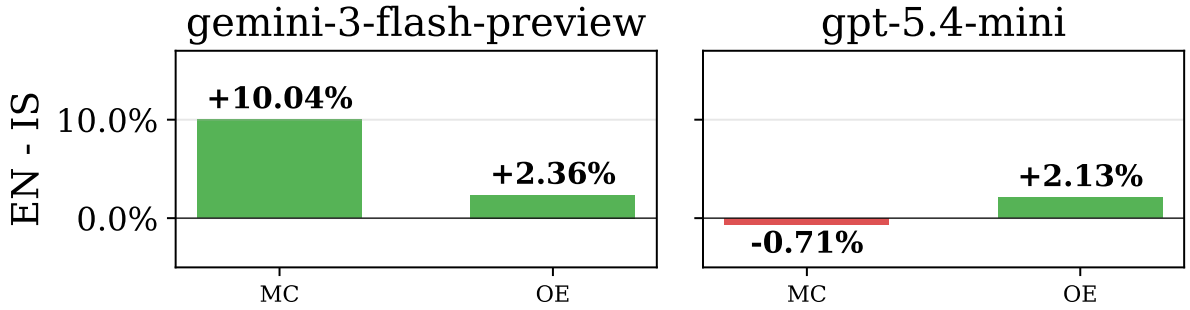

Figure 2: Language effect (EN−IS accuracy difference) for both models. Positive values indicate English advantage. Gemini-3-Flash shows substantial language effects, especially in MC mode, while GPT-5.4-mini shows minimal language differences in both modes.

42.50% (IS) and 41.79% (EN) in MC mode but reaches substantially higher accuracy of 68.83% (IS) and 70.96% (EN) in OE mode, a gap of approximately 27 percentage points. While 42% exceeds the 25% random baseline for four-choice questions, it remains surprisingly low given the model's OE performance.

Analysis of answer selection patterns reveals a systematic *option position bias*: GPT-5.4-mini over-selects option B (33.6% of responses versus 24.8% expected under uniform selection) and under-selects option D (15.8% versus 26.4% expected). Per-option accuracy drops monotonically with position: 49.7% when the correct answer is A, 49.5% for B, 42.5% for C, and only 29.9% for D, barely above random chance. This bias is consistent across both languages (English: B selected 36.6%, D selected 16.8%), confirming it is a model-level property rather than a language-specific artifact. Gemini-3-Flash exhibits a milder version of this pattern but compensates with substantially higher overall accuracy. The MC deficit observed here is robust to prompt variation: both models received identical prompts, and all API calls were routed through OpenRouter to ensure model-agnostic evaluation.

Prior work on the Open-LLM-Leaderboard found that models experience an average accuracy drop of approximately 25% when evaluated with open-style questions instead of multiple-choice (Myrzakhan et al., 2024). GPT-5.4-mini shows the reverse pattern, performing better in OE mode, suggesting that MC evaluation may systematically *disadvantage* certain models, which may be a result of some kind of misalignment. Recent multilingual evaluations confirm that language significantly influences ostensibly language-agnostic capabilities, with larger models improving average performance but not universally closing cross-lingual gaps (Huang et al., 2025).

5.4. Evaluation Robustness

The three-judge quorum addresses potential self-enhancement bias in LLM-based evaluation (Gu et al., 2024). By using judges from three independent providers (Google, OpenAI, Anthropic), we ensure that no single provider's biases can drive the results. The 97.6% unanimous agreement rate across 6,776 evaluated responses provides strong evidence that the correctness verdicts are reliable. In the 2.4% of cases where judges disagreed, the majority-vote mechanism ensures robustness.

6. Conclusion

Language does affect LLM mathematical reasoning, but the effect is model-dependent and interacts with evaluation format. Gemini-3-Flash demonstrates a consistent, significant English advantage of 2.4–10.0 percentage points, while GPT-5.4-mini shows no significant language effects in either mode. The MC format itself proves highly consequential: GPT-5.4-mini exhibits a 27 percentage point MC deficit driven by a systematic option position bias favoring earlier answer positions.

For educators and assessment designers deploying LLMs in Icelandic and other low-resource languages (Wang et al., 2024), these findings carry practical weight. Model selection requires validation on the target language, not just English benchmarks; a model that excels in English may underperform or behave unpredictably in the deployment language. Evaluation format introduces an additional variable since the same model can produce vastly different accuracy depending on whether problems are presented as MC or OE. Before classroom deployment, practitioners should test candidate models on representative problems in the target language and in the intended evaluation format, rather than extrapolating from English-only or single-format results.

A natural next step is to expand the STAK benchmark to other low-resource languages. If the model-dependent language effects we observe for Icelandic replicate across typologically diverse languages, this would strengthen the case that English-only evaluation is insufficient for deployment decisions. More broadly, competition mathematics captures only one dimension of how educators actually use LLMs. Benchmarks targeting classroom-relevant tasks, problem generation, step-by-step explanation, grading, and feedback, would give practitioners a more direct basis for model selection in their actual workflows.

7. Limitations

Several limitations should be noted. First, while we mitigate self-enhancement bias through the three-judge quorum, GPT-5.4-mini participates as both an evaluated model and a judge. The quorum design ensures that its self-judgment is checked by two independent models, and the 97.6% unanimous agreement rate suggests minimal bias impact. Second, while translations were generated using Gemini-3-Flash (one of the evaluated models), a manual review of 100 random samples by a native Icelandic speaker with professional English proficiency confirmed semantic faithfulness. Gemini-3-Flash's superior performance on English versions provides further evidence of translation quality: if translations introduced errors, we would expect degraded English performance rather than the observed improvement. Third, results are specific to the evaluated model versions and may not generalize to future releases, though the MC deficit we observe is consistent across multiple OpenAI model generations (see Appendix). Fourth, competition-level problems may not reflect typical educational use cases. Finally, the dataset represents a single low-resource language, limiting generalizability to other language families, though Icelandic's morphological complexity makes it a reasonably demanding test case for Germanic and other inflectional languages.

8. Ethical Considerations

All problems are drawn from publicly available Icelandic mathematics competitions and contain no personal data; nor do the machine translations or model outputs. LLM evaluation results for educational contexts should inform but not replace pedagogical judgement, performance on competition-style problems does not directly predict classroom suitability. Large-scale LLM inference across multiple models and conditions also carries computational and environmental costs.

9. Bibliographical References

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand. Association for Computational Linguistics.

Alexander R Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. 2025. MultiNRC: A

challenging and native multilingual reasoning evaluation benchmark for LLMs. *arXiv preprint arXiv:2507.17476*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on LLM-as-a-judge. *The Innovation*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. BenchMAX: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-LLM-Leaderboard: From multi-choice to open-style questions for LLMs evaluation, benchmark, and arena. *CoRR*.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. 2025. OLMo 3. *arXiv preprint arXiv:2512.13961*.

OpenAI. 2023. GPT-3 dataset statistics. https://github.com/openai/gpt-3/tree/master/dataset_statistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual LLM evaluation for European languages. *arXiv preprint arXiv:2410.08928*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

10. Language Resource References

Hafsteinn Einarsson and Jökull Ari Haraldsson and Ívar Armin Derayat and Sigrún Helga Lund and Benedikt Steinar Magnússon. 2026. *Icelandic Math Eval: A Competitive Mathematics Benchmark for Large Language Models*. University of Iceland.

Appendix: GPT-5.2 Results and Cross-Generational MC Deficit

The MC deficit observed in GPT-5.4-mini seems to reflect a broader pattern across OpenAI model generations as we had originally used GPT-5.2 using the same dataset and identical prompts, with correctness assessed by a single Gemini-3-Flash judge (this evaluation preceded our adoption of the three-judge quorum). The results with GPT-5.4-mini confirm that the MC deficit persists across model generations.

GPT-5.2 achieved 49.2% (IS) and 48.9% (EN) in MC mode versus 79.5% (IS) and 81.6% (EN) in OE mode, a gap of approximately 31 percentage points, compared to GPT-5.4-mini's 27-point gap. Like GPT-5.4-mini, GPT-5.2 showed no significant language effects in either mode.

Letter selection analysis reveals that GPT-5.2 exhibits a similar but distinct position bias: it over-selects options B (29.6%) and C (32.2%) while

under-selecting A (16.5%). Whereas GPT-5.4-mini's bias concentrates on option B, GPT-5.2's distributes across B and C. Both models substantially under-select compared to uniform expectation on at least one option.

The persistence of the MC deficit across two model generations, with consistent magnitude (27–31 percentage points) but shifting option preferences, suggests a systematic architectural or training characteristic of OpenAI models rather than a bug in any single release. The identical prompts and evaluation infrastructure across all experiments rule out prompt-related causes.