Exploring LLMs mathematical reasoning capability: Insights from GSM-Symbolic in English and Polish

Michał Tomczyk Marcin Łomiński

Faculty of Electrical Engineering
Warsaw University of Technology
Warsaw, Poland

Introduction

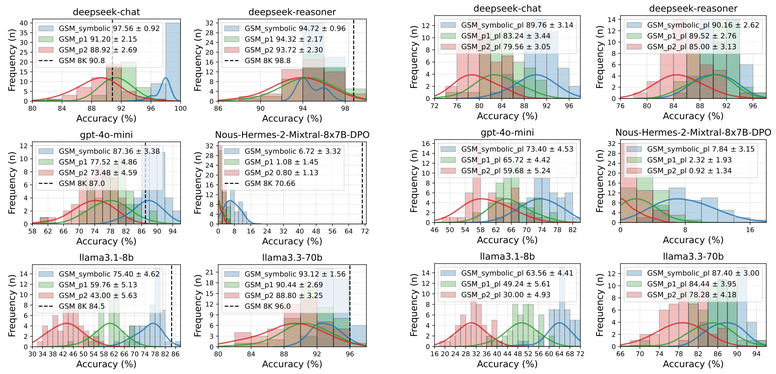
Large language models (LLMs) show progress on benchmarks like GSM8K, but it's unclear if this stems from reasoning or pattern recognition. Current benchmarks, often static and English-centric, fail to assess generalization across linguistic or logical variations. While GSM-Symbolic (Mirzadeh et al., 2024) introduced logical perturbations, language invariance — consistent reasoning across languages — remains unexplored. To bridge this gap, we present GSM-Symbolic-PL, a Polish-translated version of GSM-Symbolic, enabling bilingual evaluation of reasoning robustness. Our study systematically tests whether modern LLMs demonstrate abstract reasoning or rely on localized pattern matching, conducting large-scale experiments across multiple model families.

Research and Experiments

We evaluated state-of-the-art LLMs including `ChatGPT-4o-mini`, `DeepSeek-V3/R1`, and `LLaMA-3.3-70B` using a 3-shot Chain-of-Thought (CoT) prompting strategy. Each model completed 50 runs of 50 questions (2,500 iterations) on two datasets: the original English `GSM-Symbolic` and the Polish-translated `GSM-Symbolic-PL`. The latter was generated via Google's Translation API and manually reviewed to ensure logical equivalence. All numerical values and symbolic operators were preserved, and names remained unchanged. Only grammatical adjustments — such as verb conjugation and word order — were introduced to match Polish syntax. For example:

"Pavel is 22 years old..." became *"Pavel ma 22 lata..."*

The CoT approach provided three step-by-step examples in Polish, encouraging intermediate reasoning and mimicking human-like problem-solving. This allowed us to probe two key aspects: logical robustness (via perturbed structures) and language invariance. Although GSM8K benchmark scores (black dashed line) are shown for reference, many values are approximate due to limited transparency for some models, especially in under-researched multilingual contexts.



(a) English datasets (b) Polish datsets

Figure 1. A set of histograms comparing the performance of the 3-shot CoT method across all datasets.

Results

We present results using histograms and distribution curves, showing model accuracy across 50 runs per configuration. Mean accuracy and standard deviation are listed alongside each dataset.

Most models performed well on GSM-Symbolic. For example, ChatGPT-4o-mini achieved 87.36% ± 3.38%, but dropped to 73.40% ± 4.53% on GSM-Symbolic-PL, highlighting sensitivity to language changes—a trend visible across all models.

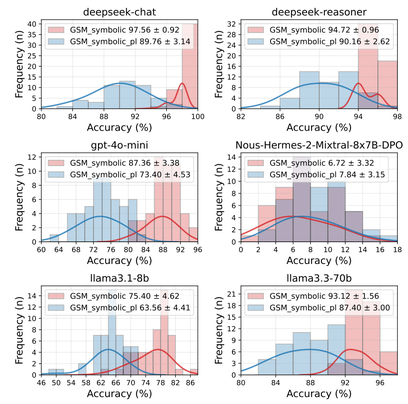
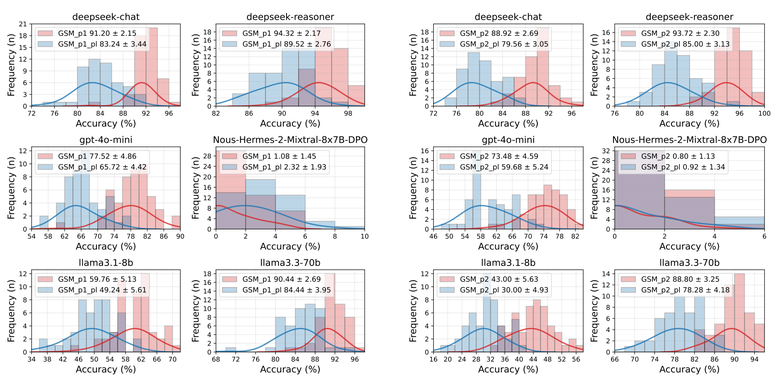


Figure 2. A set of histograms comparing the performance of GSM-Symbolic, basic dataset in both languages



(a) Performance of GSM-p1 dataset in PL and ENG (b) Performance of GSM-p2 dataset in PL and ENG

Figure 3. A set of histograms comparing the performance of GSM-P1 and GSM-P2, across both, translated datasets

Commentary

Figure 1 presents performance distributions across 50 sets for six language models on both English (GSM-Symbolic, GSM-p1, GSM-p2) and Polish (GSM-Symbolic-PL, GSM-p1-PL, GSM-p2-PL) datasets. The DeepSeek group—especially DeepSeek-Reasoner—showed strong and consistent performance across all sets (86–100%), with minimal drop in Polish. In contrast, Nous-Hermes-2-Mixtral-8x7B-DPO underperformed in all scenarios (below 7% in English, marginally better in Polish), despite a 70.66% average on GSM8K, likely due to training data overlap. Across models, a general accuracy drop in Polish highlights sensitivity to language, though the overall shape of distribution curves remains similar. Average GSM8K scores are not shown for Polish due to lack of reliable benchmarks.

Figure 2 compares GSM-Symbolic results in English and Polish. All models perform worse in Polish, except Nous-Hermes-2-Mixtral-8x7B-DPO, which on average scores marginally better in Polish. DeepSeek models are more consistent in English (94–100%), with wider spread in Polish (80–100%).

Figure 3 shows GSM-p1 and GSM-p2 results for English and Polish. Across both datasets, models consistently perform worse in Polish, except for Nous-Hermes-2-Mixtral-8x7B-DPO. However, its Polish results remain very low and likely reflect randomness rather than improved reasoning. DeepSeek models, while strong overall, show greater distribution spread here than in GSM-Symbolic, indicating reduced consistency.

Conclusion

This study assessed LLMs' mathematical reasoning using the GSM-Symbolic dataset (English & Polish) with 3-shot Chain-of-Thought. Results show significant language-based performance gaps. ChatGPT-4o-mini scored 87.36% ± 3.38% in English, dropping to 73.40% ± 4.53% in Polish, suggesting reliance on language-specific cues over abstract reasoning. DeepSeek-Reasoner remained consistent in English (86–100%) but declined in Polish. Nous-Hermes-2-Mixtral-8x7B-DPO scored under 7% in English and 0.8% on harder sets, despite a 70.66% GSM8K result—likely from training exposure. Marginally better performance in Polish datasets is rather caused by randomness than any indication of better reasoning.

Findings show CoT aids procedural steps but not generalizable logic. LLMs still lack language-agnostic reasoning. Future work should focus on benchmarks isolating reasoning from memorization and enhancing multilingual robustness.

References

[1] E. Boix-Adsera, O. Saremi, E. Abbe, S. Bengio, E. Littwin, J. Susskind, "When can transformers reason with abstract symbols?" Apple, ICLR, 2024.

[2] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio and M. Farajtabar, "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," Apple, 2024.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Google Research, Brain Team, 2023.

[4] B. Jiang, Y. Xie, Z. Hao, X. Wang, T. Mallick, W. J. Su, C. J. Taylor, D. Roth, "A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners," University of Pennsylvania, Argonne National Laboratory, 2024.

[5] OpenAI, "GPT-4 Technical Report," OpenAI, 2024.

[6] H. Zhang, J. Da, D. Lee, V. Robinson, C. Wu, W. Song, T. Zhao, P. Raja, C. Zhuang, D. Slack, Q. Lyu, S. Hendryx, R. Kaplan, M. Lunati, S. Yue, "A Careful Examination of Large Language Model Performance on Grade School Arithmetic," Scale AI, 2024.

[7] S. Mishra, D. Khashabi, C. Baral, H. Hajishirzi, "Cross-Task Generalization via Natural Language Crowdsourcing Instructions," Allen Institute for AI, University of Washington, Arizona State University, 2024.

[8] K. Valmeekam, M. Marquez, A. Olmo, S. Sreedharan, S. Kambhampati, "PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change," School of Computing & AI Arizona State University, Department of Computer Science, Colorado State University, 2023.

[9] S. Kambhampati, "Can Large Language Models Reason and Plan?," School of Computing & Augmented Intelligence, Arizona State University, 2024.

[10] DeepSeek-AI, "DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model," DeepSeek-AI, 2024

[11] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLM via Reinforcement Learning," DeepSeek-AI, 2025