Contribution ID: **39** Type: **Poster**

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

Thursday 24 April 2025 10:45 (30 minutes)

Large language models (LLMs) are trained with ever-increasing amounts of data. It seems that when asked to solve mathematical tasks, they can infer and reason mathematically [1]. The GSM8K benchmark platform is widely used to test various LLMs to solve simple arithmetic tasks. Recently, LLMs have shown a clear improvement in their ability to correctly answer questions from the GSM8K dataset. However, it is impossible to say conclusively from the GSM8K studies whether the performance increase was also followed by improvements, in the mathematical reasoning mentioned above. A recent study that shed light on the problem of mathematical reasoning in LLMs [2] attempted to address this problem by creating a database based on GSM8K but with the ability to make appropriate changes to the content of math tasks, which would test how true the claim that LLMs can reason is.

In our research, we have attempted to confirm previously established research results, but also to extend them to include newly developed language models such as DeepSeek or more advanced versions of ChatGPT. Above that, the research was also extended to check the influence of the language in which the test math tasks were written on the effectiveness of a given model. We decided to translate the GSM-Symbolic datasets using the Google Translator API, creating Polish equivalents, which we have provisionally named GSM-Symbolic-PL, GSM-P1-PL, GSM-P2-PL. The selected LLMs are then questioned hundreds of times using the 3-shot Chain-of-Thought prompting method [3], which involves giving the model 3 sample questions and answers to indicate how the model should ”think”when generating and answer to the final question. Using it is supposed to allow LLMs to be more directed towards proper mathematical thinking and reasoning which we want to investigate. After that, the results were validated and analyzed.

Our research was focused on ChatGPT versions 4o-mini and o3-mini and DeepSeek's latest versions V3 and R1. As expected, the model's responses depend on the content of the task and is strongly dependent on whether bias is placed in the task, in which case the model makes errors very often. This seems to indicate that LLMs might not be able to reason mathematically the way humans do. Their answers are inconsistent, even though the reasoning path to solve the task remains the same. GSMSymbolic, compared to GSM8K, exposed this weakness even more, but this does not change the fact that the higher models'correctness of response is quite high.

Authors: ŁOMIŃSKI, Marcin (Wydział Elektryczny, Politechnika Warszawska); TOMCZYK, Michał (Wydział Elektryczny, Politechnika Warszawska)

Presenters: ŁOMIŃSKI, Marcin (Wydział Elektryczny, Politechnika Warszawska); TOMCZYK, Michał (Wydział Elektryczny, Politechnika Warszawska)

Session Classification: Session B (Poster)