STROHMAIER, Anselm R.; VAN DOOREN, Wim; SESSLER, Kathrin;
GREER, Brian; & VERSCHAFFEL, Lieven
Ludwigsburg, Leuven, München, Portland

Word Problem Solving in Large Language Models

Theoretical Background

Mathematical word problems are challenging for Large Language Models (LLMs), since they require a combination of fundamentally different skills: Understanding a problem that is written in a natural language, finding an underlying mathematical structure and appropriate operations, performing the necessary calculations and interpreting the outcomes (Liu et al., 2023).

From a mathematics education perspective, these challenges seem not at all new: They reflect in many ways the challenges that students face when solving word problems (Verschaffel et al., 2000). Yet, mathematics education research on word problem solving has rarely considered LLMs to be a resource to better understand or teach word problem solving.

In computer sciences the term *mathematical reasoning* is commonly used as an umbrella term for solving any mathematical tasks (Sundaram et al., 2024). This differs from its use in mathematics education, where it is often associated with logical inference, argumentation, or strategic thinking (Hjelte et al., 2020). Moreover the term *word problem* can refer to very different tasks. In mathematics education, a distinction is often made between "dressed up" *prototype word problems* where language is merely an obstacle for identifying the underlying basic arithmetic operation, and *complex word problems* where factors like redundant, missing, or implicit information, real-life contexts, and the lack of a unique answer provide additional challenges and make up for the key solution processes (Verschaffel et al., 2000). In computer sciences, word problems are often referred to as elementary school, basic arithmetic problems (Lai et al., 2024). The relation between *mathematical reasoning* in computer sciences and *word problem solving* in mathematics education has not been investigated in detail so far.

In this scoping review, we summarized existing research overviews on how LLMs solve mathematical word problems, and situate this research in the mathematics education traditions. We aimed at answering the following research questions:

a) Which kinds of word problems are commonly used to evaluate the performance of LLMs?

b) How well do LLMs perform on these mathematical word problems, and how similar are solution processes to those of humans?

In: L. Schick, M. Platz & A. Lambert (Hrsg.),
Beiträge zum Mathematikunterricht 2025.
58. Jahrestagung der Gesellschaft für Didaktik der Mathematik. WTM.
https://doi.org/10.37626/GA9783959873307.0

Method

To address the rapidly evolving field, we conducted a systematic literature search, including journal articles, conference proceedings, and non-peer-reviewed preprints. We focused on summarizing papers (surveys, syntheses, or reviews) assessing LLM performance in solving mathematical word problems. Searches in *Web of Science* and *Scopus* identified 53 studies, which two authors screened ($\kappa = .79$), yielding 19 relevant abstracts. Full-text inspection based on inclusion criteria narrowed this to eight papers: three preprints, two proceedings, and three journal articles.

Results

Which kind of word problems are commonly used to evaluate the performance of LLMs?

For benchmarking LLMs, it is common to use standard datasets to compare different models and methods (Ahn et al., 2024, Saraf et al., 2024). These benchmark datasets often vary complexity with regard to the semantic structure and the number of operations and agents. One of the most commonly used datasets is GSM8K (Cobbe et al., 2021), containing 8500, 2 to 8-step arithmetic word problems created by human writers that "a bright middle school student should be able to solve“ (p.3). More difficult datasets exist, for example MATH (Hendrycks et al., 2021), which contains 12500 problems from high-school competitions from various mathematical areas (see Fig. 1). Yet, some of these problems might not be considered word problems in mathematics education due to the missing contextualization.

Because these datasets are open-source they might be included in the training data for recent LLMs. Moreover, many problems from datasets like GSM8K share a similar semantic structure, first giving an initial state and then listing equations in natural language. Similar to humans, LLMs could therefore learn the “rules of the game” of word problem solving (Verschaffel et al., 2000) by identifying typical structures, keywords, operations, procedures and answers.

Fig. 1.: Example items from GSM8K (left) and MATH (right)

Griffin had 24 french fries, but Kyle took 5 of them. Billy took twice as many as Kyle. Ginger gave Griffin a handful of her fries, and then Colby took from Griffin 3 less than the number of fries that Kyle had taken. If in the end Griffin had 27 fries, how many fries did Ginger give Griffin?	Donut Haven fries donuts in batches of 20, but sells them in boxes of 13. If Donut Haven fries just enough batches of 20 to pack 44 full boxes of 13 donuts, how many donuts will be left over?

How well do LLMs perform on these mathematical word problems, and how do solutions relate to human solution processes?

When reviewing the performance of LLMs, the most striking finding is how rapidly performance has been increasing throughout the past two years due to more complex models, larger training datasets and improved LLM architectures. Prior to 2023, LLMs were considered promising yet inferior to previous approaches of using AI for solving mathematics tasks (Ahn et al., 2024; Lu et al., 2023; Testolin, 2024). Surveys that were published throughout 2023 reported that LLMs had caught up with humans, but still struggled considerably with complex word problems and number sense (Testolin, 2024). Today, LLMs outperform students in many word problems.

For example, performance on GSM8K for GPT-3 without any specific enhancement method was reported at about 15% in 2022 (Plaat et al., 2024), and is beyond 90% for GPT-4o today (Seßler et al., 2024). For the MATH dataset, performance today is around 50%, depending on the difficulty level and methods used (Seßler et al, 2024), which is about the same as for a graduate student (Hendrycks et al., 2021). Both examples from GSM8K and MATH in Fig. 1 are consistently solved by GPT-4o today.

The accuracy of a a base model using zero-shot prompting (which means, for example, simply typing a word problem into ChatGPT), can be improved by several strategies (Ahn et al., 2024, Plaat et al., 2024). For practical reasons, changing the base model itself is not feasible for most mathematics educators, as it would require substantial technical capabilities and considerable resources to train or fine-tune an LLM. Methods that exploit the frozen base model include in-context learning, for example through advanced prompts, and the integration of external tools. Both are very similar to human strategies: Prompting can provide the model with additional information like worked examples or cognitive strategies (e.g., “try to solve the problem step by step”). External tools usually include calculators or more sophisticated coding environments that can help overcome LLMs notorious deficiencies in number sense. However, despite these similarities, prompting and external tools are often designed from a computer science perspective, not from a mathematics education perspective. For example, using knowledge about metacognition or mathematical modelling could lead to interesting insights into how transferable solution processes can be between LLMs and humans. At the same time, providing worked examples of problems with a similar structure might be helpful for LLMs and humans alike, but is problematic from an educational perspective, as it enforces the idea of superficial, routine-based word-problem solving.

Discussion

Our first glance at the literature on LLMs solving word problems shows a focus on elementary, often dressed-up problems (Lai et al., 2024). LLMs now often outperform students but face similar challenges, and strategies to improve their performance are relatable. Thus, parallels exist to mathematics education, but critical exploration of these connections is required.

References

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). Large Language Models for Mathematical Reasoning: Progresses and Challenges. In N. Falk, S. Papi, & M. Zhang, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. St. Julian's, Malta.

Cobbe, K., Kosaraju, V., Mohammad, B., Mark, C., 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.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). *Measuring Mathematical Problem Solving With the MATH Dataset.* ArXiv.

Hjelte, A., Schindler, M., & Nilsson, P. (2020). Kinds of Mathematical Reasoning Addressed in Empirical Research in Mathematics Education: A Systematic Review. *Education Sciences, 10*(10), 289.

Lai, H., Wang, B., Liu, J., He, F., Zhang, C., Liu, H., & Chen, H. (2024). *Solving Mathematical Problems Using Large Language Models: A Survey.* SSRN.

Liu, W., Hu, H., Zhou, J., Ding, Y., Li, J., Zeng, J., He, M., Chen, Q., Jiang, B., Zhou, A., & He, L. (2024). *Mathematical Language Models: A Survey.* ArXiv.

Lu, P., Qiu, L., Yu, W., Welleck, S., & Chang, K.-W. (2023). A Survey of Deep Learning for Mathematical Reasoning. In A. Rogers, J. Boyd-Graber, & N. Okazaki, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) Toronto, Canada.

Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., & Back, T. (2024). *Reasoning With Large Language Models, a Survey.* ArXiv.

Saraf, A., Kamat, P., Gite, S., Kumar, S., & Kotecha, K. (2024). Towards Robust Automated Math Problem Solving: A Survey of Statistical and Deep Learning Approaches. *Evolutionary Intelligence, 17*(5), 3113-3150.

Seßler, K., Rong, Y., Gözlüklü, E., & Kasneci, E. (2024). *Benchmarking Large Language Models for Math Reasoning Tasks.* ArXiv. https://doi.org/10.48550/arXiv.2408.10839

Sundaram, S. S., Gurajada, S., Padmanabhan, D., Abraham, S. S., & Fisichella, M. (2024). Does a language model "understand" high school math? A survey of deep learning based word problem solvers. *WIREs Data Mining and Knowledge Discovery, 14*(4), e1534.

Testolin, A. (2024). Can Neural Networks Do Arithmetic? A Survey on the Elementary Numerical Skills of State-of-the-Art Deep Learning Models. *Applied Sciences, 14*(2).

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems.* Swets & Zeitlinger.