An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP)*

Paulo Shakarian*, Abhinav Koyyalamudi, Noel Ngu and Lakshmivihari Mareedu

Arizona State University, 699 S Mill Ave, Tempe, AZ, 85281, USA

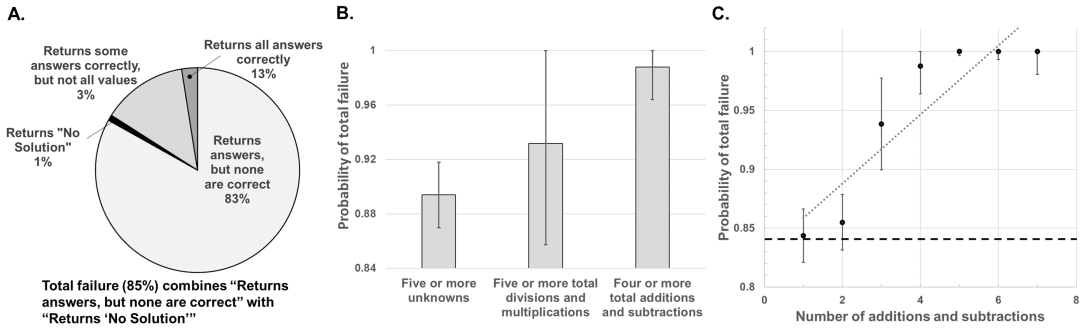

Figure 1: ChatGPT Performance on DRAW-1K math word problem (MWP) dataset. (A.) Overall results on the 1,000 MWPs in DRAW-1K based on ChatGPT's response, (B.) aspects of MWPs that led to ChatGPT failure more often than the prior (prior probability of 0.84, 95% confidence intervals shown), (C.) the increase in probability of an incorrect response as a function of the number of addition operations (prior probability shown with dashed line, 95% confidence intervals, linear regression with $R^2 = 0.821$).

Introduction. The emergence of large language models (LLM's) have gained much popularity in recent years. At the time of this writing, some consider OpenAI's GPT 3.5 series models as the state of the art [1]. In particular, a variant tuned for natural dialogue known as ChatGPT [2], released in November 2022 by OpenAI, has gathered much popular interest, gaining over one million users in a single week [3]. However, in terms of the accuracy LLMs have known performance issues, specifically when reasoning tasks are involved [1, 4]. This issue, combined with the ubiquity of such models has led to work on prompt generation and other aspects of the input [5, 6]. In other areas of machine learning, such as meta learning [7, 8] and introspection [9, 10] attempts to predict when a model will succeed or fail for a given input. An introspective tool, especially for certain tasks, could serve as a front-end to an LLM in a given application.

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023.

*You can use this document as the template for preparing your publication.

*Corresponding author.

pshak02@asu.edu (P. Shakarian); akoyyala@asu.edu (A. Koyyalamudi); nngu2@asu.edu (N. Ngu); lmareedu@asu.edu (L. Mareedu)

https://labs.engineering.asu.edu/labv2/ (P. Shakarian)

As a step toward such a tool, we investigate aspects of math word problems (MWPs) that can indicate the success or failure of ChatGPT on such problems. Of specific interest are data sets such as DRAW-1K [11, 12, 13] which not only includes 1,000 MWPs with associated answers but also template algebraic equations that one would use to solve such a word problem. For example, given the question *The student-teacher ratio for Washington High was reported to be 27.5 to 1. If there are 42 teachers, then how many students are there?* DRAW-1K includes the answer (1,115) but also template $a * m = b * c$. This information represents a symbolic representation of the problem which can potentially be used to identify aspects that make such problems more difficult. While there has been previous work examining the LLM performance on MWPs [4], such work did not investigate specific aspects that increase MWP difficulty nor did it examine performance on ChatGPT in particular.

Contributions. The contribution of this paper is as follows: (1.) the creation of a data set consisting of ChatGPT responses to the 1,000 DRAW MWPs available at **https://github.com/lab-v2/ChatGPT_MWP_eval**, (2.) we identified that ChatGPT fails in 84% of DRAW-1K problems, even if we accept partial and rounded solutions (Figure 1, panel A), (3.) we have identified several factors about MWPs relating to the number of unknowns and number of operations that lead to a higher probability of failure when compared with the prior (Figure 1, panel B), and (4.) we identified that the probability of failure increases linearly with the number of addition and subtraction operations (Figure 1, panel C). We believe that researchers studying this data set can work to develop models that can combine variables, operate directly on the symbolic template, or even identify aspects of the template from the problem itself in order to predict LLM performance. We note that at the time of this writing, collecting data at scale from ChatGPT is a barrier to such work as API's are not currently directly accessible, so this data set can facilitate such ongoing research without the overhead of data collection (note that we built our solution for feeding ChatGPT MWPs using a software available at https://github.com/mmabrouk/chatgpt-wrapper).

Related Work. The goal of this challenge data set is to develop methods to introspect a given MWP in order to identify how an LLM (in this case ChatGPT) will perform. Recent research in this area has examined MWPs can be solved by providing a step-by-step derivation [14, 15, 16, 17]. While these approaches provide insight into potential errors that can lead to incorrect results, this has not been studied in this prior work. Further, the methods of the aforementioned research are specific to the algorithmic approach. Work resulting from the use of our challenge data set could lead to solutions that are agnostic to the underlying MWP solver - as we treat ChatGPT as a black box. We also note that, if such efforts to introspect MWPs are successful, it would likely complement a line of work dealing with "chain of thought reasoning" for LLMs [5, 6] which may inform better ways to generate MWP input into an LLM (e.g., an MWP with fewer additions may be decomposed into smaller problems). While some of this work also studied LLM performance on Math Word Problems (MWPs), it only looked at how various prompting techniques could improve performance rather than underlying characteristics of the MWP that leads to degraded performance of the LLM.

Limitations. We note that DRAW-1K is notoriously difficult, with current state-of-the-art [18, 16] obtaining 59% accuracy (which still outperforms ChatGPT). We are working to create data sets for simpler MWPs [13] in addition to investigating ChatGPT's nondeterminism.

Acknowledgments

Some of the authors have been funded by the ASU Fulton Schools of Engineering.

References

[1] How does gpt obtain its ability? tracing emergent abilities of language models to their sources, URL: https://yaofu.notion.site.

[2] Chatgpt: Optimizing language models for dialogue, URL: https://openai.com/blog/chatgpt/.

[3] Chatgpt gained 1 million users in under a week. here's why the AI chatbot is primed to disrupt search as we know it. URL: https://www.yahoo.com/video/chatgpt-gained-1-million-followers-224523258.html.

[4] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. v. d. Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, L. Sifre, Training compute-optimal large language models, URL: http://arxiv.org/abs/2203.15556. arXiv:2203.15556 [cs].

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models .

[6] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, URL: http://arxiv.org/abs/2203.11171. doi:10.48550/arXiv.2203.11171. arXiv:2203.11171 [cs].

[7] T. Hospedales, A. Antoniou, P. Micaelli, A. Storkey, Meta-learning in neural networks: A survey 44 5149–5169. URL: https://www.computer.org/csdl/journal/tp/2022/09/09428530/1twaJR3AcJW. doi:10.1109/TPAMI.2021.3079209, publisher: IEEE Computer Society.

[8] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, C. C. Loy, Domain generalization: A survey 1–20. doi:10.1109/TPAMI.2022.3195549, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] S. Daftry, S. Zeng, J. A. Bagnell, M. Hebert, Introspective perception: Learning to predict failures in vision systems, URL: http://arxiv.org/abs/1607.08665. doi:10.48550/arXiv.1607.08665. arXiv:1607.08665 [cs].

[10] M. S. Ramanagopal, C. Anderson, R. Vasudevan, M. Johnson-Roberson, Failing to learn: Autonomously identifying perception failures for self-driving cars 3 3860–3867. URL: http://arxiv.org/abs/1707.00051. doi:10.1109/LRA.2018.2857402. arXiv:1707.00051 [cs].

[11] S. Upadhyay, M.-W. Chang, K.-W. Chang, W.-t. Yih, Learning from explicit and implicit supervision jointly for algebra word problems, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 297–306. URL: https://aclanthology.org/D16-1029. doi:10.18653/v1/D16-1029.

[12] S. Upadhyay, M.-W. Chang, Annotating derivations: A new evaluation strategy and dataset for algebra word problems, URL: http://arxiv.org/abs/1609.07197. doi:10.48550/arXiv.1609.07197.

[13] Y. Lan, L. Wang, Q. Zhang, Y. Lan, B. T. Dai, Y. Wang, D. Zhang, E.-P. Lim, MWPToolkit: An open-source framework for deep learning-based math word problem solvers 36 13188–13190. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21723. doi:10.1609/aaai.v36i11.21723, number: 11.

[14] Z. Gong, K. Zhou, X. Zhao, J. Sha, S. Wang, J.-R. Wen, Continual pre-training of language models for math problem understanding with syntax-aware memory network, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, pp. 5923–5933. URL: https://aclanthology.org/2022.acl-long.408. doi:10.18653/v1/2022.acl-long.408.

[15] K. S. Ki, D. Lee, B. Kim, G. Gweon, Generating equation by utilizing operators : GEO model, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, pp. 426–436. URL: https://aclanthology.org/2020.coling-main.38. doi:10.18653/v1/2020.coling-main.38.

[16] B. Kim, K. S. Ki, S. Rhim, G. Gweon, EPT-x: An expression-pointer transformer model that generates eXplanations for numbers, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, pp. 4442–4458. URL: https://aclanthology.org/2022.acl-long.305. doi:10.18653/v1/2022.acl-long.305.

[17] Y. Xia, F. Li, Q. Liu, L. Jin, Z. Zhang, X. Sun, L. Shao, ReasonFuse: Reason path driven and global–local fusion network for numerical table-text question answering 516 169–181. URL: https://www.sciencedirect.com/science/article/pii/S0925231222011444. doi:10.1016/j.neucom.2022.09.046.

[18] B. Kim, K. S. Ki, D. Lee, G. Gweon, Point to the expression: Solving algebraic word problems using the expression-pointer transformer model, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 3768–3779. URL: https://aclanthology.org/2020.emnlp-main.308. doi:10.18653/v1/2020.emnlp-main.308.