# 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*

Abstract

We study the performance of a commercially available large language model (LLM) known as ChatGPT on math word problems (MWPs) from the dataset DRAW-1K. To our knowledge, this is the first independent evaluation of ChatGPT. We found that ChatGPT's performance changes dramatically based on the requirement to show its work, failing 20% of the time when it provides work compared with 84% when it does not. Further 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, specifically noting (across all experiments) that the probability of failure increases linearly with the number of addition and subtraction operations. We also have released the dataset of ChatGPT's responses to the MWPs to support further work on the characterization of LLM performance and present baseline machine learning models to predict if ChatGPT can correctly answer an MWP. We have released a dataset comprised of ChatGPT's responses to support further research in this area.

**Keywords**

Large Language Models, Math Word Problems, ChatGPT

## 1. Introduction

The emergence of large language models (LLM) has 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 accuracy, LLMs are known to have 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]. Other areas of machine learning, such as meta-learning [7, 8] and introspection [9, 10] attempt 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.

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. We found that ChatGPT's

*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.*

*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)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

performance changes dramatically based on the requirement to show its work, failing $20\%$ of the time when it provides work compared with $84\%$ when it does not. Further several factors about MWPs can lead to a higher probability of failure when compared with the prior, specifically noting that the probability of failure increases linearly with the number of addition and subtraction operations (across all experiments). We also have released the dataset of ChatGPT's responses to the MWPs to support further work on the characterization of LLM performance. 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.

The remainder of this paper proceeds as follows. In Section 2, we describe our methodology. Then we describe our results in Section 3. Using these intuitions, we present baseline models to predict the performance of ChatGPT in Section 4. This is followed by a discussion of related work (Section 5) and future work (Section 6).

## 2. Methodology

**MWP Dataset.** In our study, we employed the DRAW-1K dataset [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. As a running example, consider the following MWP.

> *One whole number is three times a second. If 20 is added to the smaller number, the result is 6 more than the larger.*

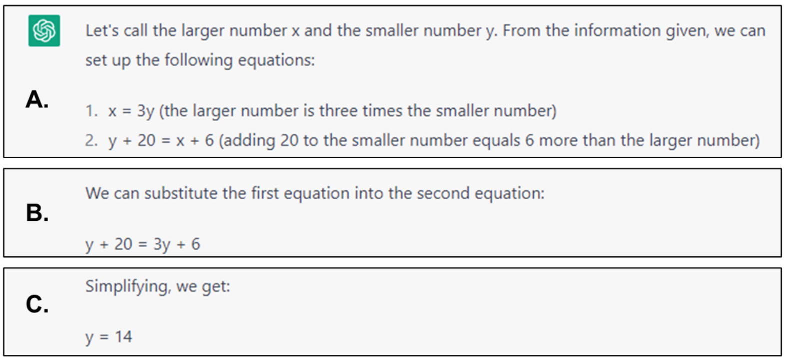
We show ChatGPT's (incorrect) response to this MWP in Figure 1. The DRAW-1K dataset not only includes the correct answer, which in this case is 12 and 7 but also includes template equations used to solve the problem. For our running example, this consists of the equations $m - n = a - b$ and $c \times m - n = 0$. This information represents a symbolic representation of the problem which can potentially be used to identify aspects that make such problems more difficult.

**Entering Problems into ChatGPT at Scale.** At the time of our study, OpenAI, the maker of ChatGPT had not released an API. However, using the ChatGPT CLI Python Wrapper[1] we interfaced with ChatGPT allowing us to enter the MWP's at scale. For the first two experiments, we would add additional phrases to force ChatGPT to show only the final answer. We developed these additions to the prompt based on queries to ChatGPT to generate the most appropriate phrase. However, we found in our third experiment that this addition impacted results. We ran multiple experiments to test ChatGPT's ability with these problems.

- **January 2023 Experiment (No work).** Our first experiment was run in early January 2023 prior to OpenAI's announcement of improved performance on mathematical tasks on January 30, 2023[2] and in this experiment we included the following statement as part of the prompt.

[1] We used ChatGPT CLI Python Wrapper by Mahmoud Mabrouk, see https://github.com/mmabrouk/chatgpt-wrapper

[2] https://help.openai.com/en/articles/6825453-chatgpt-release-notes



Let's call the larger number x and the smaller number y. From the information given, we can set up the following equations:

1. x = 3y (the larger number is three times the smaller number)
2. y + 20 = x + 6 (adding 20 to the smaller number equals 6 more than the larger number)

We can substitute the first equation into the second equation:

y + 20 = 3y + 6

Simplifying, we get:

y = 14

**Figure 1:** ChatGPT's response (Jan. 24, 2023) to MWP *One whole number is three times a second. If 20 is added to the smaller number, the result is 6 more than the larger.* In Step A it correctly identifies the set of equations needed to solve the problem and correctly simplifies it in Step B. However, it fails to correctly perform the algebraic operation in Step C (it should state $2y = 14$). This leads ChatGPT to obtain an incorrect result, returning $42$ and $14$ instead of $21$ and $7$.

> Don't provide any work/explanation or any extra text. Just provide the final number of answers for the previous question, with absolutely no other text. if there are two or more answers provide them as a comma separated list of numbers.

- **February 2023 Experiment (No work).** Our second experiment was run in mid-February 2023 after the aforementioned OpenAI announcement and also used a prompt that would cause ChatGPT to show only the answer, however we found that our original prompt led to more erratic behavior, so we modified the prompt for this experiment, and used the following.

  > Don't provide any work/explanation or any extra text. Just provide the final number of answers for the previous question, with absolutely no other text. if there are two or more answers provide them as a comma separated list of numbers like: '10, 3,' etc; or if there is only 1 answer provide it like '10'. Absolutely no other text just numbers alone. Just give me the numbers (one or more) alone. No full stops, no spaces, no words, no slashes, absolutely nothing extra except the 1 or more numbers you might have gotten as answers.

- **February 2023 Experiment (Showing Work).** We also repeated the February experiment without the additional prompt, thereby allowing ChatGPT to show all its work. We note that in this experiment we used ChatGPT Plus which allowed for faster response. At the time of this writing, ChatGPT Plus is only thought to be an improvement to accessibility and not a different model.[3]

[3] https://openai.com/blog/chatgpt-plus/

## 3. Results

The key results of this paper are as follows: (1.) the creation of a dataset consisting of ChatGPT responses to the MWPs, (2.) identification of ChatGPT failure rates (84% for January and February experiments with no work and 20% for the February experiment with work), (3.) identification of 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 3), (4.) identification that the probability of failure increases linearly with the number of addition and subtraction operations (Figure 5), and (5.) identification of a strong linear relationship between the number of multiplication and division operations and the probability of failure in the case where ChatGPT shows its work.

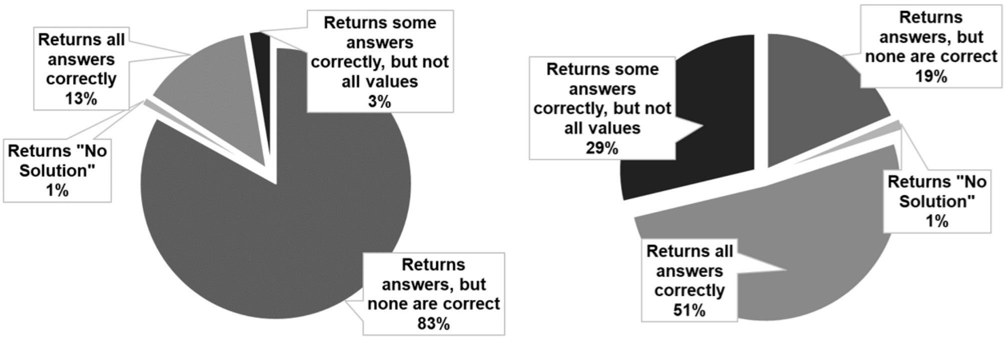
**Dataset.** We have released ChatGPT's responses to the 1,000 DRAW-1K MWP's for general use at **https://github.com/lab-v2/ChatGPT_MWP_eval**. We believe that researchers studying this dataset can work to develop models that can combine variables, operate directly on the symbolic template, or even identify 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 dataset can facilitate such ongoing research without the overhead of data collection.

**Overall Performance of ChatGPT on DRAW-1K.** As DRAW-1K provides precise can complete answers for each problem, we classified ChatGPT responses in several different ways and the percentage of responses in each case is shown in Figure 2.

1. *Returns all answers correctly.* Here ChatGPT returned all answers to the MWP (though it may round sometimes).
2. *Returns some answers correctly, but not all values.* Here the MWP called for more than one value, but ChatGPT only returned some of those values.
3. *Returns "No Solution."* Here ChatGPT claims there was no solution to the problem. This was not true for any of the problems.
4. *Returns answers, but none are correct.* Here ChatGPT returned no correct answers (e.g., see Figure 1).

Throughout this paper, we shall refer to the probability of failure as the probability of cases 3 and 4 above (considered together). In our February experiment, we found that when ChatGPT omitted work, the percentages, as reported in Figure 2 remained the same, though they differed significantly when work was included. We also report actual numbers for all experiments in Table 1. We note that the probability of failure increases significantly when the work is not shown. However, when the work is included, ChatGPT obtains performance in line with state-of-the-art models (i.e. EPT [18, 16]) which has a reported 59% accuracy while ChatGPT (when work is shown) has fully correct (or rounded) answers 51% of the time, but can be viewed as high as 80% if partially correct answers are included.

**Factors Leading to Incorrect Responses.** We studied various factors from the templated solutions provided for the MWP in the DRAW-1K dataset and these included number of equations, number of unknowns, number of division and multiplication operations, number of addition and



Performance of ChatGPT on DRAW-1K when not showing work
(same distribution for both Jan. and Feb. 2023)

Performance of ChatGPT on DRAW-1K when showing work
(Feb. 2023)

**Figure 2:** Overall results on the 1,000 MWPs in DRAW-1K based on ChatGPT's response.

| Response Type | Jan. 2023<br>(No work) | Feb. 2023<br>(No work) | Feb. 2023<br>(Showing work) |
|---|---|---|---|
| Returns answers, but none are correct | 831 | 830 | 186 |
| Returns "No Solution" | 9 | 10 | 14 |
| Returns all answers correctly | 135 | 134 | 513 |
| Returns some answers correctly, but not all values | 25 | 26 | 287 |

**Table 1**
Number of responses for each ChatGPT Variant

subtraction operations, and other variants derived from the metadata in the DRAW-1K dataset. We identified several factors that, when present, cause ChatGPT to fail with a probability greater than the prior (when considering the lower bound of a $95\%$ confidence interval). These results are shown in Figure 3. One interesting aspect we noticed is that when the system would be required to show its work, the number of unknowns present no longer seems to increase the probability of failure (this was true for all quantities of unknowns in addition to what is shown in Figure 3). Additionally, the number of multiplication and division operations, while increasing the probability of failure greater than the prior in the January experiment was not significant (based on $95\%$ confidence intervals) in the February experiment (when work was not shown) - possibly a result of OpenAI's improvements made at the end of January. However, there was a significant relationship between the number of multiplication and division operations and failure when work was shown. In fact, we found a strong linear relationship ($R^2 = 0.802$) for this relationship in the case where work was shown.

**Correlation of failure with additions and subtractions.** Previous work has remarked on the failure of LLM's in multi-step reasoning [1, 4]. In our study, we identified evidence of this phenomenon. Specifically, we found a strong linear relationship between the number of addition and subtraction operations with the probability of failure ($R^2 = 0.821$ for the January experiment, $R^2 = 0.870$ for the February experiment and $R^2 = 0.915$ when work was shown).

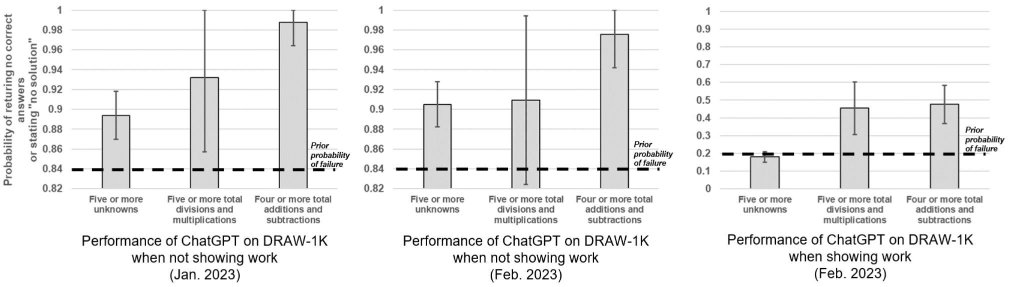


**Figure 3:** Aspects of MWPs that led to ChatGPT failure more often than the prior ($95\%$ confidence intervals shown).

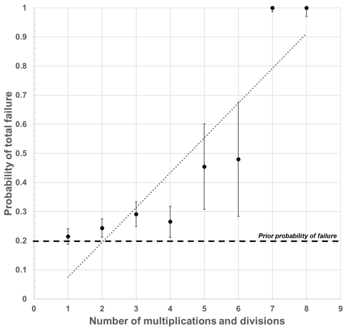


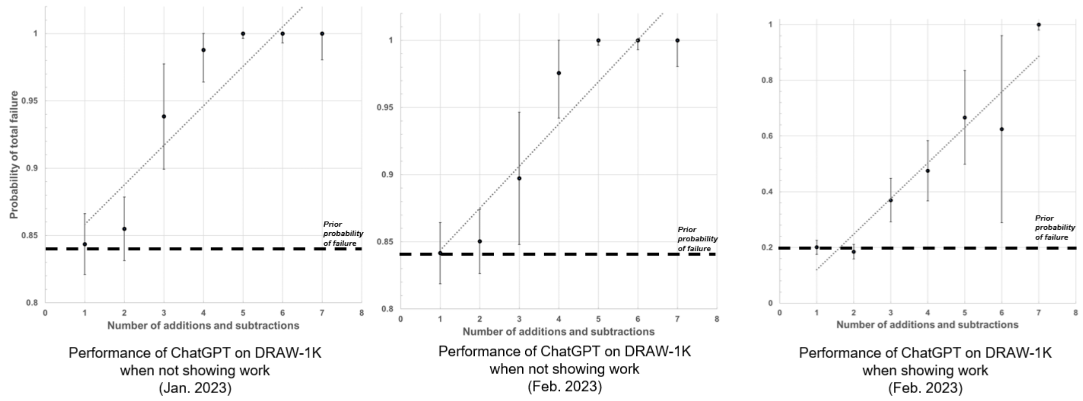
**Figure 4:** Additional finding specific to the February, 2023 experiment where ChatGPT displayed its work relating number of multiplications to probability of failure, $R^2 = 0.802$, $95\%$ confidence intervals.

We show this result in Figure 5. It is noteworthy that the relationship existed in all of our experiments, and seemed to be strengthened when ChatGPT included work in the result.

## 4. Performance Prediction Baselines

The results of the previous section, in particular, the factors indicating a greater probability of failure (e.g. Figures 3-5), may indicate that the performance of ChatGPT can be predicted. In this section, we use features obtained from the equations associated with the MWPs to predict performance. Note that here we use ground-truth equations to derive the features, so the models presented in this section are essentially using an oracle - we leave extracting such features from equations returned by ChatGPT or another tool (e.g., EPT [18]) to future work. That said, as these features deal with counts of operations, unknowns, and equations, a high degree of accuracy in creating the equations would not be required to faithfully generate such features.

Following the ideas of machine learning introspection [9, 10], we created performance prediction models using random forest and XGBoost. We utilized scikit-learn 1.0.2 and XGBoost 1.6.2 respectively. In our experiments, we evaluated each model on each dataset using a five-fold cross-validation and report average precision and recall in Table 2 (along with F1 computed based on those averages). In general, our models were able to provide higher precision than random on predicting incorrect answers for both classifiers. Further, XGBoost was shown to be



Performance of ChatGPT on DRAW-1K when not showing work (Jan. 2023)

Performance of ChatGPT on DRAW-1K when not showing work (Feb. 2023)

Performance of ChatGPT on DRAW-1K when showing work (Feb. 2023)

**Figure 5:** 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 $R^2 = 0.821$ for January, $R^2 = 0.870$ for February without showing work and $R^2 = 0.915$ for February with showing work).

| Version of | Model | Incorr. | Incorr. | Incorr. | Corr. | Corr. | Corr. |
|---|---|---|---|---|---|---|---|
| ChatGPT | Type | Prec. | Recall | F1 | Prec. | Recall | F1 |
| Jan. | RF | 0.90 | 0.88 | 0.89 | 0.34 | 0.41 | 0.37 |
| (No work) | XGBoost | 0.95 | 0.22 | 0.36 | 0.16 | 0.93 | 0.26 |
| Feb. | RF | 0.94 | 0.89 | 0.91 | 0.47 | 0.63 | 0.54 |
| (No work) | XGBoost | 0.98 | 0.35 | 0.51 | 0.18 | 0.95 | 0.31 |
| Feb. | RF | 0.78 | 0.69 | 0.73 | 0.74 | 0.82 | 0.78 |
| (Showing work) | XGBoost | 0.77 | 0.59 | 0.67 | 0.69 | 0.83 | 0.75 |

**Table 2**
Performance Prediction Baseline Models using Ground Truth Equations

able to provide high recall for predicting correct responses. While these results are likely not suitable for practical use, they do demonstrate that the features extracted provide some amount of signal to predict performance and provide a baseline for further study.

## 5. Related Work

The goal of this challenge dataset 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 dataset 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.

## 6. Future Work

Understanding the performance of commercial black-box LLMs will be an important topic as they will likely become widely used for both commercial and research purposes. Further future directions would also include an examination of ChatGPT performance on datasets other MWPs [13], investigating ChatGPT's nondeterminism, and exploring these studies on upcoming commercial LLM's to be released by companies such as Alphabet and Meta.

## 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.