PMWP: A Benchmark for Math Word Problem Solving in Persian

Marzieh Abdolmaleki, Mehrnoush Shamsfard†, Veronique Hoste and Els Lefever
LT3 – Ghent University, Belgium
†NLP Lab – Shahid Beheshti University, Iran
{marzieh.maleki,veronique.hoste,els.lefever}@ugent.be
m-shams@sbu.ac.ir

Abstract

Mathematical reasoning captures fundamental aspects of human cognitive ability. Although recent advances in LLMs have led to substantial improvements in automated mathematical problem solving, most existing benchmarks remain focused on English. As a result, robust mathematical reasoning remains a challenging and insufficiently explored capability for underrepresented languages including Persian. To address this gap, we introduce PMWP, the first dataset of 15K elementary-level Persian math word problems that supports both supervised training and evaluation of reasoning models. By expanding mathematical reasoning resources beyond English, PMWP contributes to the development of multilingual AI systems with stronger reasoning capabilities. In this work, we conduct a systematic evaluation of the Persian math word problem solving capabilities of different state-of-the-art LLMs. Our results indicate that DeepSeek-V3 exhibits reduced language bias when problem texts are translated into English, while Gemini-2.5-Flash achieves the highest equation value accuracy (72.02%) in Persian. In addition, we investigate parameter-efficient adaptation for equation generation by applying LoRA-based fine-tuning to LLaMA-3-8B and Qwen-2.5-7B. Our results show that, following fine-tuning, these open-weight models achieve 91.65% and 92.53% exact equation match accuracy, respectively. Overall, our findings provide insights into the comparative strengths and limitations of proprietary and open-weight models for mathematical reasoning in Persian.

1 Introduction

Math Word Problem (MWP) solving can be defined as a Question Answering task in which a problem description includes quantitative information alongside a related question. In many cases, particularly in algebraic word problems, the question concerns an unknown variable that must be computed using the information in the problem statement. A solver system must therefore either compute the final numerical answer directly or generate a mathematical equation that leads to the correct numerical answer. As such, MWP solving requires a combination of natural language understanding, numerical reasoning, and symbolic manipulation, making it an especially desirable capability for large language models (LLMs). An example of an MWP is shown in Table 1.

English Text:
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. How many pieces of pizza did they eat in total?
Equation: $x = 10 \times 8$ **Answer:** 80

Table 1: An English translated example of a math word problem from the PMWP dataset.

Recent models have demonstrated substantial improvements on mathematical reasoning benchmarks. GPT-4 (OpenAI, 2023) achieves over 92% accuracy on grade-school mathematics problems in GSM8K (Cobbe et al., 2021) when using chain-of-thought prompting (Wei et al., 2023), reflecting notable progress in both linguistic understanding and logical reasoning. Despite these advances, mathematical reasoning remains unevenly studied across languages. While numerous benchmark datasets have been developed to assess mathematical reasoning capabilities, most are limited to English (Patel et al., 2021; Miao et al., 2020; Koncel-Kedziorski et al., 2016). A small number of datasets have been introduced for other languages, including Chinese (Wang et al., 2017; Zhao et al., 2020; Qin et al., 2021), as well as Bangla (Prama et al., 2025), and Romanian (Cosma et al., 2025).

In contrast, despite the widespread use of LLMs by Persian speaking communities, particularly for educational purposes, mathematical reasoning in

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 74–82

Persian remains largely unexplored. The existing resources (Abaskohi et al., 2024) are designed solely for evaluation and provide no training data, limiting their usefulness for robust model development and systematic analysis.

To address this gap, we introduce the PMWP dataset, consisting of 15,588 elementary school-level MWPs annotated with their corresponding equations and numerical answers. PMWP is the first Persian benchmark with training, development, and test splits, enabling both model training and systematic evaluation. Two key characteristics of a high-quality MWP dataset are *scale*, measured by the number of problems, and *diversity*, reflected in vocabulary size. However, previous work has shown that models can often achieve high accuracy by exploiting shallow heuristics rather than genuine reasoning (Patel et al., 2021). To mitigate this issue, we apply data augmentation techniques that treat each quantitative component of a problem as a variable that can be recomputed under different conditions (Kumar et al., 2022). This design encourages models to rely less on memorized patterns and more on underlying reasoning processes. Moreover, it enables the same problem structure to be assessed from multiple perspectives, providing a more robust evaluation of the ability to reason mathematically.

Finally, we evaluated the performance of several widely used LLMs, including both commercial API-based models and open-weight models, on the PMWP dataset. This evaluation provides insights into the current capabilities and limitations of state-of-the-art LLMs for mathematical reasoning in Persian, and allows us to analyze language bias toward English in both answer prediction and equation generation settings. Our contributions are summarized as follows:

- We introduce PMWP, the first dataset for Persian MWP solving, designed for solving problems through equation generation and reasoning evaluation, with standardized training, validation, and test splits.
- We provide a systematic evaluation of popular open-weight and proprietary LLMs, including DeepSeek-V3, Gemini-2.5-Flash, and GPT-4o, on mathematical reasoning in Persian, and analyze language bias toward English for both answer prediction and equation generation.
- We study parameter-efficient adaptation of open-weight LLMs, including Llama-3-8B and Qwen-2.5-7B, using LoRA for Persian MWP solving through equation generation.
- The dataset and fine-tuned models are publicly available[1] to support future research on mathematical reasoning in Persian.

2 Related Work

Early work in this area focused on elementary-level arithmetic problems involving single-unknown equations, leading to foundational datasets such as ADDSUB (Hosseini et al., 2014), SINGLEEQ (Koncel-Kedziorski et al., 2015), and MULTIARITH (Roy and Roth, 2015). These datasets established early task formulations for mapping natural language descriptions to linear equations and laid the groundwork for subsequent MWP solvers in English.

Building on these foundations, larger and more diverse datasets were introduced to support systematic evaluation. In English, MAWPS (Koncel-Kedziorski et al., 2016) consolidated several early datasets into a unified benchmark. In Chinese, MATH23K (Wang et al., 2017) marked a significant milestone by providing over 23,000 elementary-level MWPs with equation annotations, and has since become a standard benchmark for equation generation and arithmetic reasoning. Subsequent datasets such as ASDIV (Miao et al., 2020) and SVAMP (Patel et al., 2021) emphasized problem diversity and robustness.

In parallel with dataset development, recent advances in LLMs have substantially improved performance on mathematical reasoning tasks. Techniques such as chain-of-thought prompting (Wei et al., 2023) and equation or program generation (Romera-Paredes et al., 2024; Imani et al., 2023) have enabled models such as GPT-4 (OpenAI et al., 2024) to achieve strong results on established English benchmarks. However, these evaluations remain largely English-centric, with Chinese datasets serving as the primary non-English alternative.

Prior work has explored benchmarking multilingual LLMs on Persian mathematical tasks, highlighting both the potential of large models and their limitations in handling Persian linguistic structure and numerical reasoning (Abaskohi et al., 2024). The existing Persian resources used in this benchmark consist of two small evaluation-focused

[1] https://github.com/marzieh-abdolmaleki/PMWP

datasets. One includes 50 multiple choice elementary level mathematics questions without contextual problem descriptions, while the other contains 179 questions drawn from 7th and 10th grade entrance examinations and selected translations from the MATH dataset (Hendrycks et al., 2021). Despite these efforts, Persian-focused studies largely rely on small-scale evaluation sets and do not provide publicly available datasets that support supervised training. As a result, dataset scarcity remains a primary bottleneck for systematic evaluation and model adaptation in Persian.

Translation-based approaches have been proposed as a low-cost alternative for expanding multilingual coverage, including automatically translated variants of GSM8K (Chen et al., 2024). However, prior work on Persian NLP consistently shows that naïve translation often fails to capture language-specific structures and cultural conventions that are especially important for educational tasks (Khashabi et al., 2021). These limitations motivate dataset construction strategies that combine translation with targeted linguistic validation and controlled augmentation. To the best of our knowledge, no large-scale, publicly available Persian MWP dataset currently exists that supports both training and systematic evaluation. Existing Persian resources are limited in scale and designed solely for evaluation. In contrast, the PMWP dataset combines machine translation from a large, well-established benchmark (MATH23K) (Wang et al., 2017) with structured data augmentation that systematically reassigns the unknown variable within each problem. All translations are revised by Persian native speakers and culturally adapted by modifying names, entities, and events. This design enables scalable dataset construction while preserving mathematical correctness and advancing research on Persian mathematical reasoning.

3 Persian MWP Dataset

In this section, we describe our methodology to construct the Persian MWP dataset. The goal is to collect Persian MWPs at the primary school level so that problems are solvable via linear equations with one unknown. Linear equations can only include the four arithmetic operators (addition, subtraction, multiplication, and division). In this regard, machine translation (Section 3.1) and data augmentation (Section 3.2) methods are used.

Prototype Problem
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. How many pieces of pizza did they eat in total?
Equation: $X = 10 \times 8$ Answer: 80
Transformed Problem
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. They ate 80 pieces of pizza in total.
Equation: $80 = 10 \times 8$
Candidate Problem 1
Nikoo invited X people to her birthday party. They each ate 8 pieces of pizza. They ate 80 pieces of pizza in total.
Equation: $X = 80 \div 8$ Answer: 10
Candidate Problem 2
Nikoo invited 10 people to her birthday party. They each ate X pieces of pizza. They ate 80 pieces of pizza in total.
Equation: $X = 80 \div 10$ Answer: 8
Candidate Problem 3
Nikoo invited 10 people to her birthday party. They each ate 8 pieces of pizza. They ate X pieces of pizza in total.
Equation: $X = 10 \times 8$ Answer: 80

Table 2: An example of the data augmentation method.

3.1 Machine Translation

We randomly sampled 8,000 Chinese mathematical word problems from the MATH23K dataset and translated them into Persian using the Google Translate API. We selected Chinese–Persian transfer because MATH23K is a widely used benchmark for this task, contains problem types closely aligned with our target domain, and provides substantially more data than English-based datasets such as MAWPS. The translated problems were manually reviewed and corrected when necessary by qualified expert linguists holding master's or Ph.D. degrees. All reviewers were native Persian speakers from Iran and culturally adapted the problems by modifying names, entities, and events. We also identified and corrected incorrect or inconsistent equations in the original Chinese dataset. Both the translated texts and their corresponding equations were revised during this process. After quality control, we retained 5,000 Persian math word problems for our experiments.

3.2 Data Augmentation

To extend the dataset, we produced new problems from the translated prototype problems taking the following steps:

- **Question sentence to informative sentence transformation:** Interrogative sentences beginning with terms like "what," "how much," "how many", etc., undergo handwritten rule-based transformations to incorporate an unknown symbol "x" as informative sentences. Then unknown symbols in both the text and the corresponding equation are replaced with the final answer.
- **New unknown selection:** after replacing the original unknown symbol in both the text and equation with the final answer, each existing quantity in the equation which also appears in the context is used as a new unknown candidate.
- **New equation construction for the new unknown:** the altered equation is changed so that the new unknown symbol occurs at the left side of the equation.

Because the augmentation method is driven by existing quantities in equations, it operates without errors. An example is provided in Table 2. While the data augmentation method may affect vocabulary diversity, generating different variations of the same underlying MWPs helps models avoid relying on shallow heuristics or memorizing common problem patterns. Using this approach, we produced 10,588 new Persian MWPs.

Text	
Correct	91%
Low Readability	5%
Need Correction	4%
Equation	
Correct	97%
Wrong	3%
Answer	
Correct	95%
Wrong	5%

Table 3: Validity results of PMWP.

To validate the quality of the dataset, expert linguists revised 600 randomly chosen problems with their corresponding final equations and answers. As shown in Table 3, texts, equations, and final answers in the Persian dataset yield an accuracy of 91%, 97%, and 95%, respectively. Text evaluation is divided into three categories: Correct, Low Readability, and Need Correction. A Correct text demonstrates a combination of natural language fluency and sound mathematical reasoning. On the other hand, a Low Readability text may have accurate mathematical content but is difficult to comprehend due to its sentence structure. Texts that cannot be supported by mathematical reasoning are considered for correction. In addition, a correct final answer must align with the question posed in the corresponding problem text. Similarly, a correct equation is one that accurately reflects the mathematical principles within the corresponding problem text and that yields the correct final answer.

# of Problems	15,588
# of Tokens	522,298
# of Types	4,671
# of Equation Templates	133

Table 4: Statistical information of PMWP.

	Text	**Equation**
Min. Length	6	5
Max. Length	73	16
Avg. Length	31.8	8.2

Table 5: Token-Based statistical analysis of problem text and equation components in PMWP. Equation tokens include both numbers and operators.

The novel Persian MWP collection is entitled PMWP. The statistical information of the dataset is shown in Table 4 and Table 5. The dataset is divided into train, test, and validation partitions, by the ratio of 80, 10, 10, respectively.

4 Experimental Setup

In this section, we outline the experimental setup for evaluating mathematical reasoning on the PMWP dataset. We conduct zero-shot evaluations to analyze the impact of input language and output format on model performance, followed by parameter-efficient fine-tuning experiments using LoRA to assess improvements in symbolic equation generation.

Model	AnsAcc (Persian)	AnsAcc (English)	EqValueAcc (Persian)	EqValueAcc (English)
DeepSeek-V3 (%)	**71.66**	71.31 (−0.35)	70.49	69.43 (−1.06)
Gemini-2.5-Flash (%)	69.66	**73.60** (+3.94)	72.02	71.43 (−0.59)
GPT-4o (%)	53.26	**71.31** (+18.05)	66.49	67.78 (+1.29)

Table 6: Zero-shot performance of different models on the PMWP test set. AnsAcc denotes numeric answer accuracy and EqValueAcc denotes value-based equation accuracy. Values in parentheses indicate the change after translation to English.

4.1 Zero-shot Evaluation

We evaluate the zero-shot mathematical reasoning capabilities of LLMs without any task-specific fine-tuning or in-context examples. All experiments are conducted on the PMWP test set, which consists of Persian MWPs requiring the computation of a target variable x.

To disentangle the effects of output representation and input language, we define two zero-shot evaluation settings based on the required output format: numeric answer prediction and symbolic equation generation. Within each setting, we consider both direct Persian input and translation-based English input. In the direct Persian setting, models generate final answers or equations directly from the original Persian problem. In the translation-based English setting, each Persian problem is first translated into English using the same model and then solved under an English instruction. While this setting does not isolate translation quality, it reflects realistic end-to-end usage of LLMs as multilingual problem solvers.

Zero-shot Numeric Reasoning. In the numerical setting, models are instructed to solve the problem and output the final numerical value of the target variable. The corresponding prompt template, translated into English, is shown below:

> You are given a math word problem involving the variable x, enclosed in < >.
>
> Instructions:
>
> - Explain the reasoning steps clearly and completely.
> - All explanations must appear before the final line.
> - The output must end with exactly one final line.
> - The final line must contain only the final numerical answer in the following format:
>
> x = answer
>
> - Do not add any text, symbols, whitespace, or punctuation after the final line.
>
> Problem: <problem text>
>
> Answer:

Zero-shot Symbolic Reasoning. In the symbolic setting, models are required to generate a symbolic equation that represents the solution, without numerically simplifying it. The prompt template, translated into English, used in this setting is shown below:

> You are given a math word problem involving the variable x, enclosed in < >.
>
> Instructions:
>
> - Explain the reasoning steps clearly and completely.
> - All explanations must appear before the final line.
> - The output must end with exactly one final line.
> - The final line must contain only the equation solution in the following format:
>
> x = equation
>
> - Do not numerically simplify the final equation (e.g., keep $x = 10/2$, not $x = 5$).
> - Do not add any text, symbols, whitespace, or punctuation after the final line.
>
> Problem: <problem text>
>
> Equation:

Models. We conduct zero-shot evaluations using three LLMs: DeepSeek-V3 (DeepSeek-AI et al., 2025), GPT-4o (OpenAI et al., 2024), and Gemini-2.5-Flash (Comanici et al., 2025). DeepSeek-V3 is an instruction-tuned model optimized for reasoning-intensive tasks. GPT-4o is a proprietary general-purpose model with strong multilingual and reasoning capabilities. Gemini-2.5-Flash is a latency-optimized model designed for efficient inference while maintaining competitive reasoning performance. This selection allows comparison across open and proprietary models. All models are queried via their respective APIs using a unified

prompting strategy and identical decoding configurations to ensure fair comparison. We use greedy decoding with temperature set to zero and do not provide chain-of-thought exemplars. A consistent system prompt defines the model as a mathematical problem solver.

Model predictions are compared against the gold annotations in the PMWP test set using complementary metrics. For numeric reasoning, we report answer accuracy (AnsAcc) for both Persian and translation-based English inputs. For symbolic reasoning, we report value-based equation accuracy (EqValueAcc), which measures whether the numerical solution obtained by solving the generated equation matches the gold answer. As shown in Table 6, translation-based inference generally improves numeric answer accuracy, most notably for GPT-4o, indicating a strong sensitivity to input language in zero-shot reasoning. GPT-4o is also the only model that shows an improvement in equation value accuracy after translation, whereas the performance of the other models decreases under this setting. This observation is consistent with prior findings that English translated prompts often yield better performance than Persian prompts in multilingual evaluations (Abaskohi et al., 2024). However, the improvement observed for DeepSeek-V3 after translation is minimal, suggesting that this model is less sensitive to input language than the other evaluated models. This difference may also be attributed to the models' internal Persian-to-English translation capabilities, as they are evaluated in an end-to-end setting rather than under controlled translation conditions. Nevertheless, the consistent performance gains in answer accuracy observed across models indicate a systematic bias toward English inputs. In contrast, symbolic equation generation appears to be less affected by input language choice, although it remains a more challenging task overall. For direct Persian input, DeepSeek-V3 achieves stronger performance in numeric answer accuracy, while Gemini-2.5-Flash performs better in equation value accuracy, highlighting differences in model behavior across output formats and reasoning settings.

4.2 LoRA Fine-tuning

To assess the impact of parameter-efficient adaptation on symbolic mathematical reasoning, we fine-tune autoregressive language models using Low-Rank Adaptation (LoRA) (Hu et al., 2021). In contrast to the zero-shot setting, fine-tuning is performed exclusively for symbolic equation generation, where the goal is to produce an explicit equation defining the variable x.

Models. We fine-tune two open-weight autoregressive models: Qwen2.5-7B (Yang et al., 2024) and LLaMA-3-8B (AI@Meta, 2024). Both models are instruction-capable causal language models with general reasoning abilities. We select these models to evaluate the effectiveness of LoRA-based adaptation on medium scale architectures under limited computational budgets.

We again use the PMWP dataset with predefined train, validation, and test splits. Each training instance consists of a Persian problem statement and its corresponding gold equation. The fine-tuning task is formulated as conditional generation of a symbolic equation. Given a Persian math word problem, the model is instructed to output only the equation that determines the value of x, without providing explanations or numerical simplification. All outputs are constrained to the format:

```
x = <equation>
```

This formulation directly targets structural reasoning and equation construction.

LoRA Configuration and Training Setup. We apply LoRA adapters to the query, key, value, and output projection matrices of the self-attention layers. For all experiments, we use a rank of $r = 16$, scaling factor $\alpha = 32$.

Models are fine-tuned using the AdamW optimizer for three epochs. The effective batch size is 32. Training is performed in half-precision (FP16) with gradient checkpointing enabled.

Metric	Qwen2.5-7B	LLaMA-3-8B
EqMatchAcc (%)	91.65	92.53
EqMismatch (%)	0.88	0.35
Errors (%)	7.47	7.11

Table 7: Performance comparison of Qwen2.5-7B and LLaMA-3-8B on the PMWP test set. EqMatchAcc denotes exact equation match accuracy. EqMismatch refers to cases where the generated equation differs textually from the gold equation but yields the correct numerical result. Errors indicates the percentage of incorrect or invalid generated equations.

Table 7 presents the performance of Qwen2.5-7B and LLaMA-3-8B on the PMWP test set. Both models achieve high exact equation match accuracy, exceeding 91%, indicating that parameter-

efficient fine-tuning enables strong symbolic reasoning performance on elementary-level Persian MWPs. LLaMA-3-8B slightly outperforms Qwen2.5-7B in exact equation matching, while both models exhibit comparable rates of errors.

5 Conclusion

We introduced PMWP, a Persian dataset for elementary-level math word problem solving with explicit equation annotations and standardized training, validation, and test splits. The dataset is constructed by translating problems from an established benchmark and fully validating all translated instances through expert human review, followed by structured data augmentation to increase scale while preserving mathematical correctness. Our zero-shot evaluation of open-source and proprietary LLMs reveals differing sensitivities to input language, with symbolic equation generation showing greater robustness to translation than direct answer prediction. In addition, our LoRA-based fine-tuning experiments demonstrate that open-weight models can achieve high equation generation accuracy on PMWP with a limited number of trainable parameters. Overall, this work provides a systematic assessment of current LLM capabilities for mathematical reasoning in Persian and establishes PMWP as a benchmark to support future research in multilingual and reasoning.

Limitations

PMWP focuses on elementary-level math word problems solvable with single-variable linear equations. This scoped design enables controlled evaluation of foundational mathematical reasoning in Persian, but does not cover more advanced problem types such as multi-variable reasoning, non-linear equations, geometry, or probability. As a result, the findings may not directly generalize to higher-level mathematical reasoning tasks; however, PMWP provides a solid foundation for future extensions in Persian, similar to earlier studies conducted in English.

Acknowledgments

This work was supported by the Special Research Fund of Ghent University under grant number BOF.BAF.2024.0248.01. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government – department EWI.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203, Torino, Italia. ELRA and ICCL.

AI@Meta. 2024. Llama 3 model card.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Adrian Cosma, Ana-Maria Bucur, and Emilian Radoi. 2025. RoMath: A mathematical reasoning benchmark in Romanian. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 95–111, Suzhou, China. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Vivek Kumar, Rishabh Maheshwary, and Vikram Pudi. 2022. Practice makes a solver perfect: Data augmentation for math word problem solvers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4194–4206, Seattle, United States. Association for Computational Linguistics.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2023. Gpt-4 technical report. `https://arxiv.org/abs/2303.08774`. ArXiv:2303.08774.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Tabia Tanzin Prama, Christopher M. Danforth, and Peter Dodds. 2025. BanglaMATH : A Bangla benchmark dataset for testing LLM mathematical reasoning at grades 6, 7, and 8. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 134–149, Suzhou, China. Association for Computational Linguistics.

Jinghui Qin, Xiaodan Liang, Yining Hong, Jianheng Tang, and Liang Lin. 2021. Neural-symbolic solver for math word problems with auxiliary tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5870–5881, Online. Association for Computational Linguistics.

Bernardino Romera-Paredes and 1 others. 2024. Mathematical reasoning with large language models. *Nature*, 625:468–475.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems. *Preprint*, arXiv:2009.11506.