Mathematical Reasoning Datasets for AI Modelling: An Overview

Abstract:

Mathematical reasoning is a growing area of interest for defining the cognitive abilities of Large Language Models (LLMs). Unlike conventional Natural Language Processing (NLP) tasks, mathematical reasoning uses logical structure and symbolic manipulation and involves multi-stepreasoning. The rapidly evolving LLMs have prompted researchers to create various datasets to assess and improve their reasoning. In this paper, we review datasets for mathematical reasoning, ranging from basic rule-oriented to sophisticated LLM-oriented. This paper analyses datasets such as *GSM8K*, *MATH*, *MathQA*, *Geometry3K*, *FinQA*, and the newly created*MathOdyssey* dataset. The paper presents an overview of these datasets along with search strategies, inclusion and exclusion criteria, and the selection of review literature databases. The reviewed literature indicates insufficient coverageof reasoning, biases in the datasets, and limiteddomainsfor reasoning. The findings of the reviewpoint towards the need for curated datasets, topology-based evaluations, and more cross-disciplinary reasoning domains.

Keywords: Mathematical Reasoning, Large Language *GSM8K*, *MATH*, *MathQA*, *Geometry3K*, *FinQA*, *MathOdyssey*

1 Introduction

LLMs have advanced artificial intelligence by enablingsystems to understand, reason, and solve complex problems across many domains[1].One of the most significant requirements of modelling intelligence is the ability to reason mathematically, supporting structured thinking, numerical precision, and the execution of multi-step logical inferences. Unlike most other text-based reasoning tasks, mathematical reasoning requires a higher level of abstraction and consistency across both the reasoning process and the interim steps[2]. Mathematical reasoning is one area of AI that is most dependent on benchmark datasets[3]. The first available datasets were small and simplistic,focusing on either the algebraic or the arithmetic side of mathematics. However, they have slowly adapted and evolved to enable the training of deep learning models and, most recently, large-scale LLMs. Developments in datasets such as ALG514, DRAW, and MAWPS have established benchmarks for automated mathematical problem-solving, while the more recent datasets GSM8K and MATH present both greater complexity and a more comprehensive evaluation of reasoning. There are numerous mathematical reasoning datasets. Unfortunately, despite an LLM's ability to memorise, recite, and statistically compute the correct answers to a problem, true reasoning remains an elusive benchmark. Unfortunately, because the model's ability to memorise the training set is often the only way to solve a problem, models are unable to exhibit true cognitive or logical reasoning[4]. Many researchers have noted that many

of the datasets that have yielded the best results in model evaluation were used in training the model, thereby giving the model artificially good performance. The need to capture true cognitive, logical reasoning has led to the emergence of creative sets of tasks that unequivocally capture an unlimited, borderless logic [5]. The current paper evaluates the datasets created to assess reasoning as a function of mathematics, highlighting both the evolution and characteristics of the reasoning tasks involved, as well as their limitations.

The remainder of the paper is structured as follows: Section 2 presents a methodology for the review, and Section 3 presents theresults. Section 4 presents the challenges and limitations of the current study, and Section 5 presents the conclusion

2 Methodology

The literature was retrieved from prominent scholarly databases such as IEEE Xplore[6], SpringerLink[7], ScienceDirect[8], ACM Digital Library[9]. The search strings for finding relevant included use of terms such as "mathematical reasoning datasets", "LLM benchmarks," "GSM8K review"and "math problem-solving datasets" The inclusion parameters were set to consider literature published during the duration of the most recent five years (2020 to 2026) for literature collection that contained mathematical reasoning vis-a-vis the LLMs and the use of benchmark(s) datasets. The paper was included if it presented an empirical review, analysis, or a clear description of the datasets. Studies were excluded if they primarily consisted of review-based literature or analyses without original contributions,orif they relied on small sample sizes.

3 Results

This section discusses the results of the review obtained by discussing the evolution of mathematical reasoning datasets from a primitive rule-based system to modern LLM-oriented datasets along with a comparative analysis of all the datasets

3.1 The History of Datasets Associated with Mathematical Reasoning

The evolution of datasets associated with mathematicalreasoning can be described through the identification of three distinct phases of evolution:

- The primitive phase- rule-based systems
- The intermediary phase- datasets on deep learning;
- current phase- emphasis on benchmarks for large language models (LLMs).

In the primitive phase, datasets such as ALG514[10],DRAW[11], and Dolphin[12]existed for rule-based and symbolic approaches to mathematical problems. The focus was on algebraic expressions and nonsensical or vague abstractions. The created datasets were primitive andcharacterised by a lack of scope (i.e., algebraic problems) and depth (i.e., a few problem

sets). As pointed out in [10], [11], the resulting datasets did not exceed 2,000 problems, making them unsuitable for training large neural models.

Phase two incorporated advanced learning datasets like MAWPS[13], Math23K[14], and AquA[15]. These datasets have larger scope questions with annotated answers, which were pivotal to the building of neural network-based models. Math23K has over 23,000 questions and is a popular dataset for training sequence-to-sequence models. However, these datasets cannot assess questions that require complex reasoning.

The current phase focuses on LLM-centric datasets such as GSM8K[16] and MATH. GSM8K has multi-step reasoning questions, while MATH has contest-level questions. These datasets pose a greater challenge and have become the norm for assessing LLMs, but as stated, they still struggle with high-level mathematical reasoning questions beyond a high school grade level.

3.2 Modern LLM-Oriented Datasets

Modern datasets are designed to address the limitations of earlier benchmarks by offering greater complexity, more diverse domains, and improved evaluation frameworks.

The GSM8K dataset focuses on grade-school-level arithmetic problems that require multi-step reasoning[16]. Over the years, it has become a standard benchmark for evaluating reasoning. The MATH datasetincludes problems from mathematics competitions, which include advanced topics such as algebra and number theory.

In recent years, many domain-specific datasets have also been introduced. Geometry datasets such as *Geometry3K*[17], *GeoQA*[18], and *UniGeo*[19] focus on visual and symbolic reasoning. These datasets consist of diagrams and help models to interpret both textual and visual information. Financial datasets such as *FinQA*[20] and *TabMWP*[21]support numerical reasoning in real-world contexts by combining textual and tabular data.

ScienceQA[22] combines science and mathematical reasoning with multi-modal inputs. Thesedatasets highlight the importance of integrating external knowledge and theneedfor multi-step reasoning.

The most recent contributions is the *MathOdyssey* dataset as introduced in [23]. This dataset contains 387 expert-generated problems spanning high school, university, and Olympiad levels. Unlike previous datasets, *MathOdyssey*emphasises expert curation, detailed reasoning annotations, and abalanced distribution of difficulty. This makes it a more reliable benchmark for evaluating advanced reasoning capabilities.

3.3 Comparative Analysis of Datasets

Table 1: Evolution and Comparison of Mathematical Reasoning Datasets

Dataset	Domain	Complexity	Level	Key Feature
ALG514	Algebra	Low	Early	Rule-based
MAWPS	Arithmetic	Medium	DL	Flexible dataset
Math23K	Algebra	Medium	DL	Large-scale
GSM8K	Arithmetic	Medium	LLM	Multi-step reasoning
MATH	Advanced	High	LLM	Competition problems
Geometry3K	Geometry	High	LLM	Visual reasoning
FinQA	Finance	High	LLM	Real-world reasoning
ScienceQA	Science	High	LLM	Multi-modal reasoning
MathOdyssey	Mixed	Very High	LLM	Expert-curated

The analysis of the datasets, as presented in Table 1, shows a clear progression from simple rule-based datasets to complex, multi-domain benchmarks. Modern datasets emphasise diversity, interpretability, and real-world applicability.The assessment shows that datasets have a huge impact on the reasoning developed in LLM's. Towards the start, datasets concentrated mostly on symbolic reasoning. In the present day, however, datasets emphasise multi-step reasoning, practicality, and clarity. Regardless, there are many other sides of reasoning that may go unevaluated by a single dataset. The creation of expert-curated datasets, such as MathOdyssey, is a major advancement in improving the quality of evaluation. These datasets are less likely to have data contamination. More challenging datasets point to the need for evaluation frameworks specific to that field. This is the case with domain-centric datasets such as FinQA and Geometry3K. An additional avenue is the structural and topological analysis of LLM embedding spaces. Applying this, in conjunction with dataset evaluation, would yield a deeper understanding of reasoning.

4 Challenges and Limitations

Although there has been significant progress in mathematical reasoning for LLMs, several challenges remainin the dataset and the research based on it. One of the challenges of dataset bias is that models learn patterns specific to the dataset rather than develop general reasoning abilities. Another issue is that data contamination further complicates evaluation, as overlaps between training and benchmark datasets can inflate performance estimates. In addition, most datasets focus primarily on final answer correctness, with limited attention to intermediate reasoning steps, which makes it difficult to assess true understanding. Additionally, the limited availability of multilingual datasets also puts a restriction on evaluation across diverse languages and contexts.

This study has a few limitations that can be worked upon in the future. Firstly, The study is based on a qualitative analysis of selected datasets and recent literature, which may not capture the full diversity of available benchmarks. The reliance on published studies introduces potential bias, as findings depend on the quality and scope of existing work. Furthermore, the rapidly evolving nature of LLMs and dataset development means that new datasets and evaluation techniques may quickly supersede current observations, requiring continuous updates to maintain relevance.

5 Conclusion and Future Scope :

This paper presents an overview of the reasoning capabilities of Large Language Models (LLMs) baed on existing mathematical datasets.The study shows a clear progression from simple rule-based datasets to complex, multi-domain benchmarks. Modern datasets emphasise diversity, interpretability, and real-world applicability.The assessment shows that datasets have a huge impact on the reasoning developed in LLM's. Towards the start, datasets concentrated mostly on symbolic reasoning. In the present day, however, datasets emphasise multi-step reasoning, practicality, and clarity. The creation of expert-curated datasets, such as MathOdyssey, is a major advancement in improving the quality of evaluation. These datasets are less likely to have data contamination. The study also presents the challenges of these datasets and the research based on these. Also, the limitations and future scope in this direction have been discussed. It is suggested that future research should develop new datasets to address existing problems. The AI community would benefit from more comprehensive, diverse datasets that cover a wider range of languages and domains. More datasets should have more detailed annotations about reasoning to facilitate explainable AI. The community would also benefit from integrating topology-based evaluation methods that inform us about the structural characteristics of representations. Moreover, the research community needs to focus more on adaptive, dynamic datasets that evolve with the model'scapabilities.

References:

[1] Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*(10), 260.
[2] Yan, Y., Su, J., He, J., Fu, F., Zheng, X., Lyu, Y., ... & Hu, X. (2025, July). A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 11798-11827).
[3] Lu, P., Qiu, L., Yu, W., Welleck, S., & Chang, K. W. (2023, July). A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 14605-14631).
[4] Dai, W. Z., Xu, Q., Yu, Y., & Zhou, Z. H. (2019). Bridging machine learning and logical reasoning by abductive learning. *Advances in Neural Information Processing Systems*, *32*.
[5] Ghosh, P. (2025). A Critical Analysis of the Proposed Recursive Logic Subsystem for Self-Learning LLMs in Scientific Discovery.
[6] https://ieeexplore.ieee.org/Xplore/home.jsp. Last accessed on 22 March 2026.
[7] https://link.springer.com/. Last accessed on 22 March 2026
[8] https://www.sciencedirect.com/. Last accessed on 23 March 2026
[9] https://dl.acm.org/. Last accessed on 23 March 2026
[10]Huang, D., Shi, S., Lin, C. Y., Yin, J., & Ma, W. Y. (2016, August). How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 887-896).
[11] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., ... & Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
[12]Naik, A., Liu, J., Wang, C., Sethi, A., Dutta, S., Naik, M., & Wong, E. (2024). Dolphin: A programmable framework for scalable neurosymbolic learning. *arXiv preprint arXiv:2410.03348*.
[13]Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., &Hajishirzi, H. (2016, June). 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* (pp. 1152-1157).
[14] Zhao, W., Shang, M., Liu, Y., Wang, L., & Liu, J. (2020). Ape210k: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506*.
[15] Jen, T. Y., Huang, H. H., & Chen, H. H. (2021, December). Recycling numeracy data augmentation with symbolic verification for math word problem solving. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 653-657).
[16] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
[17] Ning, M., Wang, Q. F., Huang, K., & Huang, X. (2023, October). A symbolic characters aware model for solving geometry problems. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 7767-7775).
[18 Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E., & Lin, L. (2021, August). Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 513-523).
[19 Yi, X., Huang, J., Cui, F. Q., Tong, A., Wang, R., Liu, L., & Guo, D. (2026). UniGeo: A Unified 3D Indoor Object Detection Framework Integrating Geometry-Aware Learning and Dynamic Channel Gating. *arXiv preprint arXiv:2601.22616*.
[20] Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., ... & Wang, W. Y. (2021, November). Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3697-3711).

[21] Liu, R., Wang, W., Zhang, C., & Yao, Y. (2025, June). Invertible TabMap: An Invertible Self-supervised Mapping for Imbalanced Classification of Tabular Data. In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE.
[22] Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., & Bhattacharyya, P. (2022). Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, *23*(3), 289-301.
[23] Fang, M., Wan, X., Lu, F., Xing, F., & Zou, K. (2025). Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *Scientific data*, *12*(1), 1392.