Mining Misconceptions in Mathematics

Bryan Constantine Sadihin **Hector Rodriguez Rodriguez** **Matteo Jiahao Chen**
Department of Computer Science
Tsinghua University
{wangwd24,lad24,chenjiah24}@mails.tsinghua.edu.cn

1 Introduction

Multiple-choice questions are widely used to evaluate student knowledge. This work proposes a model to predict the misconceptions associated with incorrect math answers. Well-designed MCQs need a set of correct answers and incorrect answers, commonly known as distractors. Distractors align with common misconceptions and serve as tools to identify knowledge gaps. However, determining the relationship between distractors and misconceptions is a hard and time-consuming task. Automatically performing the misconception mining would increase the efficiency of the multiple-choice question assessment.

There are two major challenges in understanding misconceptions. Firstly, high-quality questions include such a wide variety of misconceptions that they cannot be conventionally classified. Secondly, large language models (LLMs) struggle to understand misconceptions because they are optimized to produce correct answers and do not reason in the same manner as humans when answering questions.

We propose the use of an optimized LLM to analyze the incorrect answers and determine the associated misconception. Afterwards, the vector embedding of the LLM answer can be compared to the vector embedding of the established misconception categories to produce a list of suitable misconceptions.

2 Definition

Evaluation Metric. We evaluate the model using the Mean Average Precision at 25 (MAP@25), a popular metric in information retrieval. It is calculated as follows:

$$MAP@25 = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{\min(n,25)} P(k) \times rel(k) \tag{1}$$

where U is the total number of observations, $P(k)$ is the precision at cutoff k, n is the number of predictions per observation, and $rel(k)$ is an indicator function. The indicator function is 1 if the item at rank k is the correct label, 0 otherwise. Once the correct label has been found, the subsequent predictions are disregarded.

3 Related work

3.1 Large Language Models for Math World Problems

Math Word Problems (MWPs) are mathematical exercises presented as written descriptions rather than direct equations [1]. LLMs must understand the relevant mathematical reasoning before formulating equations to solve the given problem. MWPs can be divided into three categories:

- **Question-Answer**: Each math question is only paired with the answer. SAT-Math [2] provides a high-school SAT Math dataset with multiple-choice questions.

Machine Learning (80245013-0) Proposal 2024.

- **Question-Equation-Answer**: Each math question is paired with the answer and the solving equation. The datasets target elementary school-level difficulty, and they are available in English (SVAMP[3], ParaMAWPS [4]) and in Chinese (Math23K [5], CM17K [6]).
- **Question-Rationale-Answer**: Each math question is paired with the answer and the reasoning path, akin to the Chain-of-Thought method, which includes the reasoning steps for correct problem-solving guidance [7]. Rationale data is generated using program induction (AQUA [8]) or LLMs (SAT-Math-COT [9])

LLMs such as Phi-3.5 [10] have demonstrated good performance in zero-shot MWP inferences. Additionally, fine-tuning an LLM with a math-specific dataset can improve the reasoning capabilities [11]. In-context learning with advanced prompting has shown promising results at a lower cost [1]. For example, an advanced prompting technique called "Self-Consistency" [12] significantly improved Chain-of-Thought by selecting the most consistent answer out of multiple LLM reasoning paths.

3.2 Vector embeddings

Vector embeddings are used in natural language processing for information retrieval [13]. One of the most popular embedding search alternatives is dense retrieval, which compares the similarities between the embedded texts to select information. The most common similarity metrics are the Euclidean distance, the cosine similarity, and the dot product similarity. Since the Euclidean distance and the dot product similarity are sensitive to both the magnitude and the direction, they can be useful for comparing embeddings that include measures. On the contrary, the cosine similarity is independent of the magnitude and can be used to compare the overall context [14].

Recent advancements in vector embeddings for information retrieval are not only limited to dense retrieval. BGE-M3 Embedding [15] uses an integrated relevance score that combines dense retrieval with sparse and multi-vector retrieval.

4 Proposed methods

Our main dataset is the Kaggle Eedi misconception dataset [16]. This dataset includes 1,868 English MWPs. Each question has one correct answer and three distractors that are aligned with one of the 2,586 possible misconceptions.

Baseline LLMs such as Phi-3.5 have a strong zero-shot capability for mathematical reasoning. Therefore, they provide a straightforward baseline using zero-shot inference to propose incorrect answer misconceptions. However, the context window of the LLM may not fit the Eedi misconception list. To address this, we will narrow down the misconception list fed to the LLM by using misconceptions' vector embedding search to retrieve correlated misconceptions.

In-context learning and fine-tuning Our proposal focuses on extending in-context learning and fine-tuning approaches for LLM mathematical reasoning to understand mathematical misconceptions. We aim to evaluate the reasoning ability difference between the in-context learning and the fine-tuning approach.

Dataset generation and external datasets Some misconceptions rarely appear in the Eedi dataset distractors. We propose using other LLMs to balance the main dataset by generating misconception-aligned distractors. Additionally, we propose to extend the dataset using external Question-Rationale-Answer data to fine-tune the LLM for accurate distractor construction using answer reasoning.

Multi-vector embedding retrieval There is a high degree of similarity between the misconceptions that characterize the answers to the same question. Thus, we propose the use of multi-vector embedding retrieval to leverage the similarity of the incorrect answers and their respective misconceptions. We will evaluate BGE-M3 performance to compute the vector embeddings of the misconception proposed by the LLM and the list of possible misconceptions. Additionally, other vector embedding techniques and similarity metrics could be explored.

References

[1] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024. EACL 2024 Student Research Workshop.

[2] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

[3] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems?, 2021.

[4] Syed Rifat Raiyan, Md Nafis Faiyaz, Shah Md. Jawad Kabir, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. Math word problem solving by generating linguistic variants of problem statements. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 362–378, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Zihao Zhou, Maizhen Ning, Qiufeng Wang, Jie Yao, Wei Wang, Xiaowei Huang, and Kaizhu Huang. Learning by analogy: Diverse questions generation in math word problem, 2023.

[6] Jinghui Qin, Xiaodan Liang, Yining Hong, Jianheng Tang, and Liang Lin. Neural-symbolic solver for math word problems with auxiliary tasks, 2021.

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

[8] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation : Learning to solve and explain algebraic word problems, 2017.

[9] Nathan Davidson. Sat math chain-of-thought dataset. `https://huggingface.co/datasets/ndavidson/sat-math-chain-of-thought`, 2023. Accessed: 2024-10-26.

[10] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, and Yi-Ling Chen. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[11] Yixin Liu, Avi Singh, C. Daniel Freeman, John D. Co-Reyes, and Peter J. Liu. Improving large language model fine-tuning for solving math problems, 2023.

[12] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[13] Jose Camacho-Collados and Mohammad Taher Pilehvar. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, 2018.

[14] Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, pages 887–890, 2024.

[15] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[16] Kaggle and Eedi. Eedi - mining misconceptions in mathematics dataset. `https://www.kaggle.com/competitions/eedi-mining-misconceptions-in-mathematics`, 2020. Accessed: 2024-10-26.