Eedi: Mining Misconceptions in Mathematics

Padmanabh Butala
pb8176@rit.edu

Harshal Chalke
hc4292@rit.edu

***Abstract*—Understanding math misconceptions can be challenging for students, often leading to wrong answers in diagnostic questions. Fixing these misconceptions involves matching wrong answers (distractors) with specific labels, which takes a lot of time and is often inconsistent when done by humans. This project focuses on creating a Natural Language Processing (NLP) model using Machine Learning (ML) to predict the connection between misconceptions and distractors in multiple-choice questions. Using data from Eedi, Vanderbilt University, and The Learning Agency Lab, our goal is to make the labeling process faster, more accurate, and reliable. The model will be tested using the Mean Average Precision @ 25 (MAP@25) score to ensure it works well for both common and new misconceptions. This solution will help teachers tag misconceptions more easily, improving the learning experience for students and making education better for everyone.**

I. Introduction

Misconceptions in mathematics often hinder students from achieving a deeper understanding of key concepts. These misconceptions are frequently reflected in the incorrect answers, or distractors, chosen by students in diagnostic assessments. Tagging these distractors with the corresponding misconceptions can provide valuable insights to educators, enabling them to tailor their teaching strategies. However, manually tagging distractors is not only time-consuming but also prone to inconsistencies. To address this issue, we are motivated to develop automated solutions that can streamline the process while maintaining high accuracy.

In this project, we approach the problem using two distinct methodologies: classical machine learning techniques and large language models (LLMs). Each methodology offers unique strengths and challenges. The classical ML methods focus on interpretable and computationally efficient algorithms, while the LLM approach leverages the power of state-of-the-art models to handle complex semantic relationships.

The classical machine learning approach consists of two key methods. The first method involves Latent Semantic Analysis (LSA), which is a statistical technique used to identify patterns in the relationships between terms and concepts in a text corpus. By reducing the dimensionality of the data, LSA enables the model to uncover latent structures that correlate misconceptions with distractors effectively. The second method utilizes word embeddings, where each word or phrase is mapped to a vector in a continuous space. These embeddings capture semantic meaning and relationships, providing a robust foundation for understanding the connections between distractors and misconceptions.

For the LLM approach, we utilize the QWEN 2.5 Instruct model, a cutting-edge language model with 7 billion parameters. This model is designed to excel in instructional tasks, which makes it particularly suited to understand and predict the nuanced relationships between misconceptions and distractors. The model's ability to process both textual and mathematical content allows it to generalize well to a wide range of questions and misconceptions. By fine-tuning this model on our dataset, we aim to achieve high performance in predicting correct misconception labels.

II. Dataset Overview

The dataset used for this project was provided by Eedi, in collaboration with Vanderbilt University and The Learning Agency Lab. It consists of diagnostic multiple-choice questions designed to assess students' understanding of various mathematical constructs. Each question includes one correct answer and three distractors, which are carefully crafted to align with specific misconceptions. This allows the dataset to serve as a valuable resource for studying the relationship between distractors and misconceptions.

Each question in the dataset is associated with a unique identifier, along with metadata such as the subject, construct, and question text. The dataset also includes labels that map each distractor to its corresponding misconception, enabling supervised learning approaches. The text content of the questions and answers was extracted from images using a human-in-the-loop OCR process, ensuring high-quality text data for analysis.

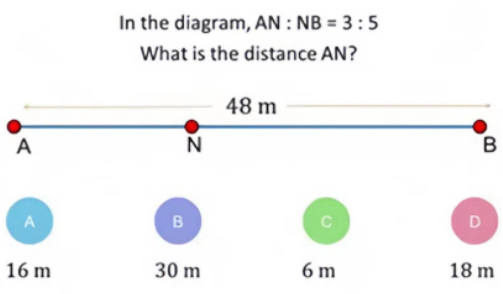

Fig. 1. Example of a diagnostic question with labeled misconceptions.

An example of a diagnostic question from the dataset is shown in Figure 1. The options are labeled with specific misconceptions as follows:

- **A:** Divides total amount by each side of the ratio instead of dividing by the sum of the parts

- **B:** Mixes up sides of a ratio
- **C:** Finds one part of a ratio but doesn't multiply that by the number of parts needed
- **D:** Correct answer

The training dataset provides both question-answer pairs and their respective misconception labels, while the test dataset requires participants to predict these labels. Additionally, the dataset includes a mapping file that links each misconception ID to its description, providing further context for model development. Evaluation of predictions is performed using the Mean Average Precision @ 25 (MAP@25) metric, which rewards models for ranking relevant misconceptions higher.

This dataset presents unique challenges due to the complexity of mathematical language and the variability in misconceptions. Furthermore, the presence of new or previously unidentified misconceptions in the test data highlights the importance of developing robust models capable of generalizing beyond the training data. These challenges underscore the significance of leveraging both classical ML methods and LLMs as separate approaches to tackle this problem effectively.

III. Methodologies

The methodologies employed in this project can be broadly categorized into three approaches: Latent Semantic Analysis (LSA), Word Embedding-Based Methods, and Large Language Models (LLMs). Each approach leverages distinct techniques to address the problem of predicting misconceptions linked to distractors in mathematical diagnostics.

A. Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) uses statistical techniques to identify relationships between words in a corpus by mapping them into a lower-dimensional semantic space [3]. Key steps in the LSA pipeline include:

- **TF-IDF Vectorization:** The *Term Frequency-Inverse Document Frequency (TF-IDF)* technique assigns weights to words based on importance, reducing the influence of common words while emphasizing rare terms [4].
- **Truncated Singular Value Decomposition (SVD):** SVD reduces the dimensionality of the TF-IDF matrix, extracting latent features that capture essential relationships between words and concepts [1].
- **Cosine Similarity:** After dimensionality reduction, cosine similarity measures the semantic closeness between text representations, such as question-answer pairs and misconception descriptions [4].

This method is computationally efficient, interpretable, and scalable for large datasets, though it may struggle to capture deeper semantic nuances.

B. Word Embedding-Based Approach

Word embeddings represent words as dense vectors in a continuous vector space, capturing semantic and syntactic relationships [7]. The embedding-based pipeline consists of:

- **Pre-trained Word Embeddings:** Models such as GloVe, FastText, and Word2Vec provide vectors for a vast vocabulary, mapping semantically similar words closer together in high-dimensional space [8].
- **Sentence Embeddings:** Sentence embeddings are generated by averaging the word embeddings of all words in a sentence, producing fixed-length representations for variable-length text [7].
- **Dimensionality Reduction:** Truncated SVD reduces the dimensionality of sentence embeddings, improving computational efficiency and revealing latent patterns [1].
- **Cosine Similarity:** Cosine similarity measures the closeness of question-answer pairs and misconceptions in the reduced space [4].

This approach captures richer semantic information, making it effective for nuanced tasks, though it may require extensive computational resources for large datasets.

C. Large Language Models (LLMs)

Large Language Models (LLMs) leverage billions of parameters to understand and process complex semantic relationships in text. In this project, the QWEN 2.5 Instruct model was employed using a two-stage pipeline:

- **Initial Retrieval with Bi-Encoder:** The *sentence-transformers/all-mpnet-base-v2* model was used to embed queries and misconceptions. Cosine similarity ranked the top 200 potential misconceptions for each query [2].
- **Re-Ranking with Cross-Encoder:** The *cross-encoder/ms-marco-MiniLM-L-12-v2* model scored query-misconception pairs, selecting the top 25 misconceptions for final predictions [5].

While the LLM-based approach offers high precision and adaptability, it requires significant computational resources and further tuning to achieve optimal performance [6].

D. Comparison of Approaches

- **LSA:** Provides a computationally efficient and interpretable approach but lacks the ability to capture deeper semantic nuances [9].
- **Word Embeddings:** Captures richer semantic relationships and performs better than LSA but requires substantial resources for large datasets [10].
- **LLMs:** Combines the efficiency of bi-encoders with the precision of cross-encoders, offering superior performance for nuanced tasks. However, it is resource-intensive and may require hybrid approaches for further improvement [5].

E. Common Preprocessing and Evaluation

All approaches shared common preprocessing and evaluation steps:

- **Data Preprocessing:** Structured tagging was applied to input text for consistent contextual representation. Tags such as `[SUBJECT]`, `[CONSTRUCT]`, and `[CHOSEN_WRONG_ANSWER]` highlighted critical information, with wrong answers emphasized using markers

(<<>>). Additional cleaning removed LaTeX markers and normalized text formatting [2].

- **Evaluation Metric:** The Mean Average Precision @ 25 (MAP@25) was used to evaluate model performance, quantifying how well the top 25 predicted misconceptions aligned with the ground truth. This metric ensured a fair comparison across methods [6].

IV. Results

The following table summarizes the MAP@25 scores for different methods:

TABLE I
EVALUATION RESULTS FOR LSA, WORD EMBEDDING, AND LLM METHODS

Method	MAP@25 Score
LSA (50 components)	0.3220
LSA (100 components)	0.3176
LSA (200 components)	0.3163
LSA (300 components)	0.3194
Best LSA (50 components)	**0.3220**
GloVe Embedding	0.3328
FastText Embedding	0.3292
Word2Vec Embedding	0.3246
Combined Embeddings	**0.3358**
Bi-Encoder (Training)	0.1169
Cross-Encoder (Validation)	0.1045

A. Steps in the Evaluation

1) **MAP@25:**

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

2) **Predicted Misconception IDs:** For each validation sample, the model outputs the top 25 misconception IDs based on cosine similarity between the reduced embeddings (e.g., LSA components, word embeddings, or bi-encoder embeddings) of the test QA pairs and the misconception mappings.
3) **Actual Misconception IDs:** Each incorrect QA pair in the validation or test set has a ground truth misconception ID provided in the dataset.
4) **Relevance Calculation:** Relevance is calculated as a binary array indicating whether the ground truth misconception ID appears in the top 25 predicted misconception IDs. The relevance is represented as a binary array where: 1 indicates that the true misconception ID is in the predictions. 0 indicates that it is not.

B. Analysis of Results

The results indicate that the Word Embedding-Based Approach achieves the highest MAP@25 score of 0.3358 when combining embeddings from GloVe, FastText, and Word2Vec. This demonstrates its ability to capture rich semantic representations effectively, outperforming LSA methods [10].

For the Large Language Models (LLMs), the bi-encoder achieved a MAP@25 score of 0.1169 on the training set, while the cross-encoder scored 0.1045 on the validation set. Although these scores are lower than the Word Embedding-Based Approach, they reflect the inherent complexity of the task and the early-stage development of the LLM-based methodology [5], [2].

C. Key Insights

- The LSA method, while interpretable and computationally efficient, struggles to capture deeper semantic relationships, leading to a maximum MAP@25 score of 0.3220 [9].
- The Word Embedding-Based Approach outperforms LSA, particularly when combining multiple embeddings, emphasizing its strength in representing nuanced textual relationships [10].
- The LLM-based pipeline shows promise in leveraging advanced models for contextual understanding but requires further tuning to improve generalization and performance. Its structured input format and multi-stage retrieval pipeline highlight a path toward scalable and robust solutions for complex tasks [5], [2].

V. Discussion and Future Directions

The evaluation of the proposed methods provides valuable insights into their performance and limitations. Each method demonstrates distinct strengths, making them suitable for different aspects of the problem.

A. Insights from Evaluation

The LSA method, though computationally efficient and interpretable, achieves its best MAP@25 score of 0.3220 with 50 components. This highlights its reliance on parameter tuning and its inability to fully leverage contextual richness due to its statistical nature [9].

The Word Embedding-Based Approach surpasses LSA, particularly with combined embeddings from GloVe, FastText, and Word2Vec, achieving the highest MAP@25 score of 0.3358. This underscores the power of pre-trained embeddings in capturing nuanced semantic relationships [10]. However, the approach demands substantial computational resources, limiting scalability for very large datasets.

The LLM-based pipeline demonstrates promising potential, with a MAP@25 score of 0.1169 (training) and 0.1045 (validation). While these scores are modest, the two-stage retrieval process—combining bi-encoder efficiency with cross-encoder precision—highlights a path toward scalable and robust solutions for addressing complex semantic tasks [5], [2]. Structured tagging and input formatting significantly contribute to its contextual understanding.

B. Challenges and Limitations

The relatively modest scores across all methods reflect the inherent difficulty of the task. Misconceptions in mathematics often involve subtle and context-dependent distinctions, posing challenges for even advanced models. The LLM approach, while powerful, is computationally intensive and requires further fine-tuning to generalize effectively [5].

C. Future Directions

The findings point to several promising avenues for future work:

- **Hybrid Approaches:** Combining the interpretability of LSA with the semantic richness of embeddings or integrating LLMs with classical methods could balance efficiency and accuracy [6].
- **Domain-Specific Models:** Fine-tuning pre-trained models on domain-specific data or developing custom embeddings tailored to misconceptions in mathematics may enhance performance [9].
- **Data Augmentation:** Incorporating additional labeled data or employing synthetic data generation techniques could help models better generalize to unseen misconceptions [2].
- **Exploring Scalability:** Optimizing computational requirements for LLMs, such as model distillation or hardware-efficient implementations, can make these methods more accessible for large-scale applications [10].

VI. Conclusion

In summary, the results highlight the potential of advanced NLP methods to address educational challenges. While significant progress has been made, further refinements are essential to achieve robust and scalable solutions for tagging misconceptions, ultimately aiding educators in improving learning outcomes.

This study demonstrates that different methodologies bring distinct strengths to the task:

- LSA provides a computationally efficient and interpretable baseline, suitable for quick analysis of text relationships.
- The Word Embedding-Based Approach excels in capturing nuanced semantic relationships, achieving the best overall performance.
- The LLM-based pipeline, while computationally demanding, shows promise with its structured and context-aware approach to misconception tagging [2], [5].

The findings emphasize the need for hybrid approaches, domain-specific fine-tuning, and efficient scaling strategies to improve performance further [6], [9]. By leveraging the strengths of each method and addressing their limitations, future work can develop more accurate and generalizable solutions. These advancements will enhance educators' ability to address student misconceptions, making personalized learning more effective and accessible.

VII. References

References

[1] Jerome R. Bellegarda. A comparison of dimensionality reduction techniques for text retrieval. *IEEE Transactions on Signal Processing*, 48(3):1012–1024, 2000.

[2] Jaekeol Choi et al. Improving bi-encoder document ranking models with two rankers and multi-teacher distillation. *arXiv preprint arXiv:2103.06523*, 2021.

[3] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman. An introduction to latent semantic analysis. In *Handbook of Latent Semantic Analysis*, pages 259–284. Psychology Press, 1990.

[4] Susan T. Dumais. Improving retrieval performance by relevance feedback using latent semantic indexing (lsi). In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 88–95. Springer-Verlag, 1994.

[5] Omar Khattab et al. In defense of cross-encoders for zero-shot retrieval. *arXiv preprint arXiv:2212.06121*, 2022.

[6] Jie Lei et al. Loopitr: Combining dual and cross encoder architectures for image-text retrieval. *arXiv preprint arXiv:2203.05465*, 2022.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[9] Iknoor Singh et al. Multistage bicross encoder: Team gate entry for mlia multilingual semantic search task 2. *arXiv preprint arXiv:2101.03013*, 2021.

[10] Wei Xiong et al. Asymmetric bi-encoder for image–text retrieval. *Multimedia Systems*, 2023.