CS 584 PROJECT : MINING MISCONCEPTIONS IN MATHEMATICS

Aadit Harshal Baldha
abaldha@stevens.edu

Harsha Vardhan Dasari
hdasari@stevens.edu

Rishi Chhabra
rchhabra1@stevens.edu

ABSTRACT

This project addresses the challenge of identifying and analyzing mathematical misconceptions in multiple-choice questions. Using natural language processing (NLP) and machine learning techniques, we aim to classify incorrect answer options based on the type of misconception they represent and evaluate how closely these misconceptions relate to common errors in mathematical understanding. The project has significant implications for educational assessment, enabling exam creators to design more effective tests and helping educators identify and address common conceptual gaps. Our results demonstrate improved methods for detecting and categorizing mathematical misconceptions compared to previous approaches used in the EEDI Kaggle competition.

1 Introduction

Mathematical misconceptions play a crucial role in educational assessment and learning. The ability to identify and classify these misconceptions is essential for creating effective examination materials and improving educational outcomes. This project focuses on analyzing multiple-choice questions to determine the relationship between incorrect answer options and common mathematical misconceptions.

The problem specifically involves determining the ”affinity” of incorrect answers to particular misconceptions. For example, when analyzing a calculus question about derivatives, some incorrect options may closely relate to common differentiation errors (high-affinity misconceptions), while others might be more distantly related to the concept (low-affinity distractors).

This challenge is important because examinations are a fundamental tool for evaluating students' understanding across various educational contexts. Creating appropriate multiple-choice questions requires careful consideration of the misconceptions that commonly arise. By developing models that can automatically identify these misconceptions, we can help educators create more effective assessment tools and better understand students' learning challenges.

The scope of this project includes applying NLP and machine learning techniques to identify and classify mathematical misconceptions in multiple-choice questions. Our objectives are to:

- Develop models that accurately detect and classify mathematical misconceptions
- Compare the effectiveness of fine-tuned large language models (LLMs) against other approaches
- Evaluate different training methods and metrics for this specialized task
- Assess the generalizability of the trained models across different mathematical domains

2 Related Work

The analysis of mathematical misconceptions using computational methods represents an emerging field at the intersection of education research and machine learning. Previous work in this area has primarily focused on rule-based approaches or simple classification methods, which often fail to capture the nuanced nature of mathematical errors.

The Kaggle competition ”EEDI - Mining Misconceptions in Mathematics” has catalyzed research in this domain, with participants employing various techniques to address the challenge. These approaches included:

- Data augmentation strategies to enhance model training on limited datasets
- Topic modeling to identify patterns in mathematical question content
- Named Entity Recognition (NER) to extract mathematical concepts and expressions
- Complex model architectures combining retrievers, rerankers, and large language models

However, these methods faced several limitations. The inconsistent labeling of misconceptions in the training data posed challenges for supervised learning approaches. Additionally, the domain-specific nature of mathematical language made it difficult for general-purpose NLP models to process the content effectively. The scarcity of labeled data for mathematical misconceptions further complicated the development of robust models.

Our work builds upon these previous approaches while addressing their limitations through improved preprocessing techniques, specialized modeling for mathematical content, and more effective evaluation metrics.

3 Methodology

To address the challenge of mining mathematical misconceptions, we formulate the problem as a multi-class classification task. For each incorrect answer option in a multiple-choice question, we aim to identify the specific misconception it represents.

Formally, given a question q with its correct answer a_c and a set of incorrect options $A_i = \{a_1, a_2, ..., a_n\}$, our goal is to learn a function $f : (q, a_i) \rightarrow m_j$ that maps each question-answer pair to a misconception label m_j from the set of possible misconceptions $M = \{m_1, m_2, ..., m_k\}$.

Our methodology consists of several key components:

3.1 Data Preprocessing

We employ specialized preprocessing techniques for mathematical text:

- Tokenization that preserves mathematical expressions and symbols
- Stemming and lemmatization adapted for mathematical terminology
- Data augmentation through paraphrasing and symbolic substitution

3.2 Feature Extraction

We extract features from both questions and answer options:

- Mathematical entity recognition to identify formulas, expressions, and numerical values
- Contextual embeddings that capture the semantic relationships between mathematical concepts
- Structural features that represent the logical organization of mathematical arguments

3.3 Model Architecture

We explore multiple modeling strategies for classifying mathematical misconceptions in multiple-choice questions, integrating both transformer-based encoders and ensemble techniques.

Fine-tuned Transformer Models

Let $\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}$ be the set of question-option pairs (queries), and $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$ be the set of possible misconceptions. We fine-tune transformer models $f_\theta(\cdot)$ (e.g., BERT, RoBERTa) to generate contextual embeddings:

$$\mathbf{q}_i = f_\theta(q_i), \quad \mathbf{m}_j = f_\theta(m_j)$$

These embeddings serve as the basis for similarity-based retrieval or classification.

Dual Encoder (Retrieval)

In the **Dual Encoder** setup, we independently encode the query and misconception texts using two transformer encoders f_q and f_m. The similarity score is computed as:

$$\text{Sim}(q_i, m_j) = \cos\left(f_q(q_i), f_m(m_j)\right)$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. For each query q_i, we retrieve the top-k misconceptions:

$$\text{TopK}(q_i) = \arg\max_{m_j \in \mathcal{M}} \text{Sim}(q_i, m_j)$$

Cross Encoder (Reranking)

The **Cross Encoder** jointly encodes the concatenated query-misconception pair (q_i, m_j):

$$s_{ij} = f_\phi\left(\text{[CLS] } q_i \text{ [SEP] } m_j\right)$$

where f_ϕ is the cross encoder transformer, and s_{ij} is the predicted relevance score. We then apply a softmax to normalize the scores:

$$P(m_j \mid q_i) = \frac{e^{s_{ij}}}{\sum_{m_l \in \text{TopK}(q_i)} e^{s_{il}}}$$

This reranking stage captures deeper contextual interactions for improved accuracy.

Ensemble Model

To improve robustness, we combine the outputs from multiple models:

- TF-IDF similarity baseline
- Dual Encoder retrieval scores
- Cross Encoder reranking scores

For each query-misconception pair (q_i, m_j), we compute the final score using a simple average of individual model scores:

$$\text{FinalScore}(q_i, m_j) = \frac{1}{3}\left(\text{TFIDF}_{ij} + \text{Sim}_{ij} + s_{ij}\right)$$

The top predicted misconceptions are selected by ranking based on this aggregated score.

Loss Function

To train our models effectively for the task of misconception classification, we utilize a contrastive loss function in the Dual Encoder setup and a cross-entropy loss in the Cross Encoder setup.

Dual Encoder Loss (Contrastive Loss): In the Dual Encoder, we apply contrastive learning to bring query-misconception embeddings closer for positive pairs and push them apart for negative ones. The loss is computed using the InfoNCE formulation:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{q}_i, \mathbf{m}_i^+))}{\exp(\text{sim}(\mathbf{q}_i, \mathbf{m}_i^+)) + \sum_{j=1}^{K} \exp(\text{sim}(\mathbf{q}_i, \mathbf{m}_{ij}^-))}$$

where:

- $\mathbf{q}_i$ is the embedding of the i-th query

- $\mathbf{m}_i^+$ is the embedding of the correct (positive) misconception
- $\mathbf{m}_{ij}^-$ are embeddings of K negative misconceptions
- $\text{sim}(\cdot, \cdot)$ denotes cosine similarity

Cross Encoder Loss (Cross-Entropy): For the Cross Encoder, we treat the reranking as a classification problem and apply the categorical cross-entropy loss:

$$\mathcal{L}_{\text{cross}} = -\sum_{i=1}^{N}\sum_{j=1}^{k} y_{ij} \log \hat{y}_{ij}$$

where:

- $y_{ij} \in \{0, 1\}$ is the ground truth label indicating whether misconception m_j is relevant for query q_i
- $\hat{y}_{ij}$ is the predicted probability (from the softmax over reranker scores)

These loss functions are optimized independently for each stage of the model pipeline.

3.4 Misconception Analysis

We develop techniques to analyze the relationships between incorrect answers and mathematical misconceptions:

- Similarity metrics to measure the affinity between answer options and common misconceptions
- Clustering methods to identify patterns in misconception distribution
- Interpretability techniques to explain model predictions

4 Experimental Setup

Our experimental framework is designed to systematically evaluate the performance of our methodology and answer key research questions about mathematical misconception mining.

4.1 Data

We utilize the dataset provided by the Kaggle competition "EEDI - Mining Misconceptions in Mathematics." This dataset includes:

- Multiple-choice mathematics questions with unique identifiers
- Construct and subject identifiers representing the mathematical topics
- Answer options labeled with corresponding misconception IDs
- Mapping between misconception IDs and descriptive names

The dataset contains information organized into several fields:

- QuestionId: Unique identifier for each question
- ConstructId and ConstructName: Specific mathematical concept being tested
- SubjectId and SubjectName: Broader mathematical topic
- QuestionText: OCR-extracted text from question images
- Answer[A/B/C/D]: Description of each answer option
- Misconception[A/B/C/D]Id: Target labels identifying the misconception represented by each option

We may perform additional data augmentation as needed to address class imbalance or enhance model training.

4.2 Evaluation Metrics

To evaluate our models, we employ **Mean Average Precision at K (MAP@K)**, a standard ranking metric particularly suited for multi-label prediction tasks such as misconception classification.

Mean Average Precision at K (MAP@K)

Mean Average Precision at K is defined as the mean of the average precision scores for each query, considering only the top K retrieved predictions.

Let Q be the set of all queries (question-option pairs), and for each query $q_i \in Q$, let AP@K$_i$ denote the average precision at K. Then the MAP@K is given by:

$$\text{MAP@K} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \text{AP@K}_i$$

The average precision at K for a single query is computed as:

$$\text{AP@K}_i = \frac{1}{\min(K, R_i)} \sum_{k=1}^{K} P_i(k) \cdot \text{rel}_i(k)$$

where:

- R_i is the number of relevant (true) misconceptions for query q_i
- $P_i(k)$ is the precision at rank k
- $\text{rel}_i(k)$ is an indicator function: 1 if the item at rank k is a relevant misconception, 0 otherwise

In our setup, we use $K = 25$, hence reporting **MAP@25** to assess how well the top-25 predicted misconceptions match the true labels.

4.3 Comparison Methods

To evaluate the effectiveness of our proposed retrieval + reranking pipeline, we compare it against the following baselines and ablations:

- **TF-IDF Baseline:** A traditional vector-space retrieval model that computes cosine similarity between TF-IDF representations of queries and misconceptions.
- **Standard Transformer Baseline:** A pretrained BERT model fine-tuned for classification without any mathematical adaptation or architectural modification.
- **Dual Encoder Only:** A model that retrieves top-k misconceptions based on independently encoded embeddings of queries and misconceptions.
- **Cross Encoder Only:** A reranker that jointly encodes query-misconception pairs to assign relevance scores.
- **Dual + Cross Encoder Pipeline:** The full retrieval followed by reranking pipeline.
- **Ensemble Model:** A simple average of scores from TF-IDF, Dual Encoder, and Cross Encoder to aggregate strengths.

To ensure robust evaluation, we applied:

- Stratified train-validation splits to maintain label distribution
- MAP@25 as the primary evaluation metric
- Additional metrics such as top-k accuracy for diagnostic purposes
- Manual inspection of error cases to assess qualitative behavior

4.4 Results

The table below summarizes the performance of different models in terms of Mean Average Precision at 25 (MAP@25). Each model is evaluated on the Testing Dataset.

Table 1: Model Performance Comparison (MAP@25)

Model	MAP@25
TF-IDF Baseline	0.350
Dual Encoder (Retrieval Only)	0.420
Cross Encoder (Reranking)	0.470
Dual + Cross Encoder (Pipeline)	0.510
Ensemble (TF-IDF + Dual + Cross)	0.530

5 Conclusion and Future Work

Our experiments show that a retrieval + reranking architecture significantly outperforms traditional baselines in identifying mathematical misconceptions. The final ensemble achieved a MAP@25 of 0.53, placing it close to top solutions in the Kaggle challenge.

Future directions include:

- **Preprocessing Variants:** Investigating the impact of different preprocessing strategies (e.g., formula normalization, symbolic abstraction, equation parsing) on model performance.
- **Error Analysis:** Conducting detailed analysis of model errors to identify misconception types that are consistently misclassified and designing targeted solutions.
- **LLM-enhanced Reranking:** Incorporating instruction-tuned or domain-adapted large language models (LLMs) for more nuanced reranking with lower inference cost.
- **Model Robustness:** Evaluating generalizability across unseen mathematical domains and noisy or OCR-imperfect input.
- **Minimal Supervision:** Exploring weak supervision or few-shot learning techniques to reduce reliance on large labeled datasets.

References

[1] Kaggle, *EEDI - Mining Misconceptions in Mathematics*, `https://www.kaggle.com/competitions/eedi-mining-misconceptions-in-mathematics`

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Proceedings of NAACL.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692.

[4] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv preprint arXiv:1908.10084.

[5] Humeau, S., Shuster, K., Lachaux, M. A., & Weston, J. (2020). *Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring*. In ICLR.

[6] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*. In ICML.