# Distractor Labeling for Improved Educational Outcomes

## Introduction

When learning mathematical foundations, students can harbor misconceptions that lead them to incorrect answers. However, identifying the misconception causing the mistake can be difficult and expensive to identify. For example, when presented the following question:

Which of the following is the simplified equivalent to $\frac{2}{3} \times \frac{3}{4}$ ?  a) $\frac{5}{7}$  b) $\frac{6}{12}$  c) $\mathbf{\frac{1}{2}}$

a) This incorrect answer indicates a misunderstanding of addition instead of multiplication.
b) This incorrect answer indicates a misunderstanding of simplification

In this project, we created a distractor labeling system leveraging the reasoning ability of Large Language Models (LLMs) to improve the understanding and management of misconceptions, enhancing the educational experience for teachers and students.

## Dataset and Analysis

We use the Kaggle dataset Eedi to develop our model, description can be found in our proposal.

**Constructs:** The dataset shows an imbalanced distribution, with common constructs like "square of a number" dominating while other constructs appear only once. Rare constructs, though infrequent, offer insights into less explored topics. A cumulative percentage analysis reveals that a small number of constructs account for most occurrences, indicating potential bias toward specific concepts.

**Subjects:** Common subjects such as "linear equations" dominate the dataset, while niche topics occur rarely, reflecting a focus on foundational areas. Subjects appearing only once highlight coverage of specialized topics. The cumulative percentage analysis confirms that a few subjects account for the majority of data, emphasizing their importance for analysis and modeling.

**Misconceptions:** Missing or incomplete Misconception IDs reveal data gaps. Common misconceptions, such as "Uses same operation instead of inverse," dominate, while rare ones reflect specific misunderstandings. A cumulative percentage analysis shows that most data is concentrated in a few misconceptions, suggesting areas for targeted interventions.

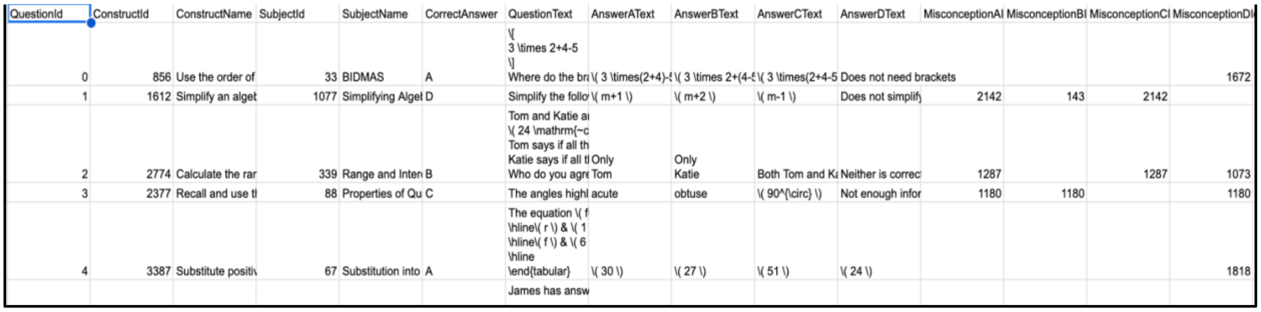
## Data Cleaning

Once the dataset had been evaluated, we needed to address the challenges posed by the missing values in the dataset, given their prevalence and the complexities of imputation.

**Missing Values:** Missing values are present only in the training set labels, specifically in the misconception fields (MisconceptionAId, MisconceptionBId, MisconceptionCId, MisconceptionDId). All rows contain at least one missing value, occurring in two cases: when the correct answer lacks a misconception (e.g., A is correct, so MisconceptionAId is NaN), or when multiple misconceptions are missing. Approximately 49% of rows fall into the latter category, making their removal detrimental to model training.

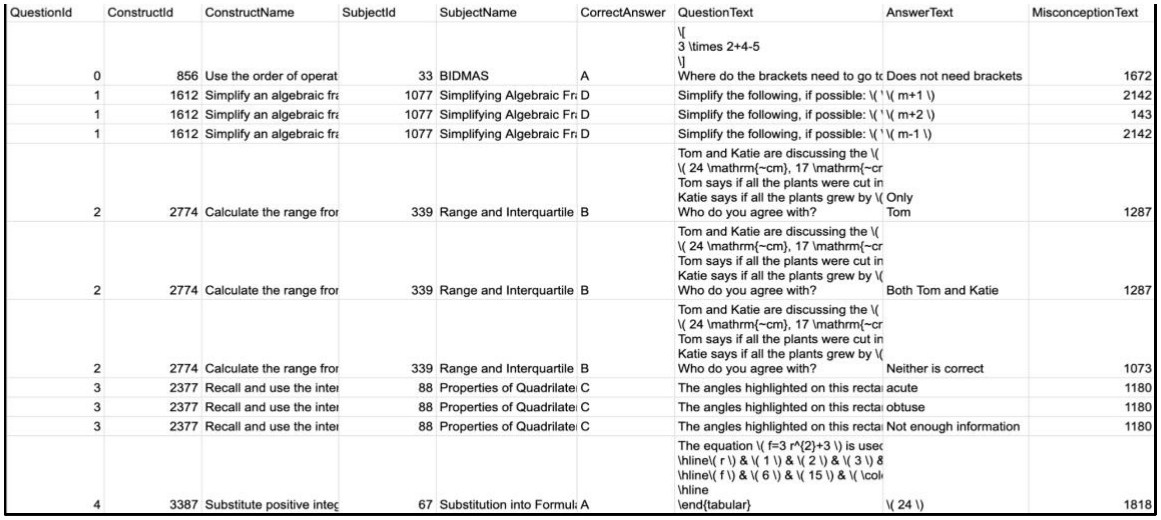
**Imputation Challenges:** Imputing missing values would require manually referencing over 2,000 misconceptions to assign appropriate values, introducing significant cost, potential for human error, and bias. This process is impractical and would compromise data reliability.

**Initial Dataset:**



| QuestionId | ConstructId | ConstructName | SubjectId | SubjectName | CorrectAnswer | QuestionText | AnswerAText | AnswerBText | AnswerCText | AnswerDText | MisconceptionAI | MisconceptionBI | MisconceptionCI | MisconceptionDI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 856 | Use the order of | 33 | BIDMAS | A | \[<br>3 \times 2+4-5<br>\]<br>Where do the bra | \( 3 \times(2+4)- | \( 3 \times 2+(4- | \( 3 \times(2+4-5 | Does not need brackets | | | | 1672 |
| 1 | 1612 | Simplify an algeb | 1077 | Simplifying Algel | D | Simplify the follo | \( m+1 \) | \( m+2 \) | \( m-1 \) | Does not simplify | 2142 | 143 | 2142 | |
| 2 | 2774 | Calculate the rar | 339 | Range and Inter | B | Tom and Katie ar<br>\( 24 \mathrm{~c<br>Tom says if all th<br>Katie says if all t<br>Who do you agre | Only<br>Tom | Only<br>Katie | Both Tom and Ka | Neither is correc | 1287 | | 1287 | 1073 |
| 3 | 2377 | Recall and use tl | 88 | Properties of Qu | C | The angles highl | acute | obtuse | \( 90^{\circ} \) | Not enough infor | 1180 | 1180 | | 1180 |
| 4 | 3387 | Substitute positiv | 67 | Substitution into | A | The equation \( f<br>\hline\( r \) & \( 1<br>\hline\( f \) & \( 6<br>\hline<br>\end{tabular} | \( 30 \) | \( 27 \) | \( 51 \) | \( 24 \) | | | | 1818 |
| | | | | | | James has answ | | | | | | | | |

**Cleaned and transformed dataset:**



| QuestionId | ConstructId | ConstructName | SubjectId | SubjectName | CorrectAnswer | QuestionText | AnswerText | MisconceptionText |
|---|---|---|---|---|---|---|---|---|
| 0 | 856 | Use the order of operat | 33 | BIDMAS | A | \[<br>3 \times 2+4-5<br>\]<br>Where do the brackets need to go tc | Does not need brackets | 1672 |
| 1 | 1612 | Simplify an algebraic fra | 1077 | Simplifying Algebraic Fr: | D | Simplify the following, if possible: \( ' | \( m+1 \) | 2142 |
| 1 | 1612 | Simplify an algebraic fra | 1077 | Simplifying Algebraic Fr: | D | Simplify the following, if possible: \( ' | \( m+2 \) | 143 |
| 1 | 1612 | Simplify an algebraic fra | 1077 | Simplifying Algebraic Fr: | D | Simplify the following, if possible: \( ' | \( m-1 \) | 2142 |
| 2 | 2774 | Calculate the range fror | 339 | Range and Interquartile | B | Tom and Katie are discussing the \(<br>\( 24 \mathrm{~cm}, 17 \mathrm{~cr<br>Tom says if all the plants were cut in<br>Katie says if all the plants grew by \(<br>Who do you agree with? | Only<br>Tom | 1287 |
| 2 | 2774 | Calculate the range fror | 339 | Range and Interquartile | B | Tom and Katie are discussing the \(<br>\( 24 \mathrm{~cm}, 17 \mathrm{~cr<br>Tom says if all the plants were cut in<br>Katie says if all the plants grew by \(<br>Who do you agree with? | Both Tom and Katie | 1287 |
| 2 | 2774 | Calculate the range fror | 339 | Range and Interquartile | B | Tom and Katie are discussing the \(<br>\( 24 \mathrm{~cm}, 17 \mathrm{~cr<br>Tom says if all the plants were cut in<br>Katie says if all the plants grew by \(<br>Who do you agree with? | Neither is correct | 1073 |
| 3 | 2377 | Recall and use the inter | 88 | Properties of Quadrilatei | C | The angles highlighted on this rectai | acute | 1180 |
| 3 | 2377 | Recall and use the inter | 88 | Properties of Quadrilatei | C | The angles highlighted on this rectai | obtuse | 1180 |
| 3 | 2377 | Recall and use the inter | 88 | Properties of Quadrilatei | C | The angles highlighted on this rectai | Not enough information | 1180 |
| 4 | 3387 | Substitute positive integ | 67 | Substitution into Formul: | A | The equation \( f=3 r^{2}+3 \) is usec<br>\hline\( r \) & \( 1 \) & \( 2 \) & \( 3 \) &<br>\hline\( f \) & \( 6 \) & \( 15 \) & \( \col<br>\hline<br>\end{tabular} | \( 24 \) | 1818 |

## Data Sampling

We opted to sample by question, ensuring that individual math questions do not appear in both training and validation sets to avoid data leakage. While this approach prevents overlap, it does not address label imbalances in misconceptions. A random 80/20 split was applied to the dataset, followed by preprocessing for training and testing.

---

## Method

We employed existing open-source lightweight pre-trained LLMs for the prediction of a misconception given question content and chosen distraction (incorrect answer option). These models are expected to possess strong reasoning capability suitable for this task, given their training on internet-scale texts. The overall framework is composed of four main components:

1. Distractors and Misconceptions Embedding: we embed distractors and misconceptions into feature space with LLM to capture the general semantic and logical relationships between them.
2. Embedding-based misconception retrieval: we use cosine similarity to retrieve top-k related misconceptions for each distractor.
3. Embedding Tuning: since the pre-trained LLM is not specialized in processing mathematical questions, we propose to train a single-layer projector to further optimize the primitive embedding.
4. Pose-retrieval Misconception Selection: a more powerful LLM is employed to re-rank the top k related misconceptions to select the most-relevant one.

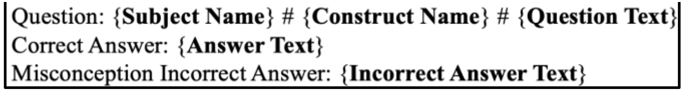
Finally, we employ Mean Average Precision@25 to evaluate our model,

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

where U is the number of observations (each observation stands for a pair of question and distraction), P(k) is the precision at cutoff k, n is the number of predictions submitted per observation, and $rel(k)$ is an indicator function equaling 1 if the item at rank $k$ is a relevant (correct) label, 0 otherwise. This metric allows our model to submit up to 25 relevant Misconceptions for each Distractor but also requires the true misconception to be ranked as high as possible.

## Distractors and Misconceptions Embedding

In this Part, we use Mistral-7B to embed the text descriptions of Distractors and Misconceptions into feature space to capture the latent semantic relationship between them. We employ the following prompt to better extract the characteristics of the Distractor, and the Misconceptions embedding is directly obtained based on its text description.



Question: {**Subject Name**} # {**Construct Name**} # {**Question Text**}
Correct Answer: {**Answer Text**}
Misconception Incorrect Answer: {**Incorrect Answer Text**}

## Embedding-based misconception retrieval

With the embedding of Distractors ***D*** and Misconceptions ***M***, we can directly retrieve the top 25 relevant Misconceptions of each Distractor by computing the cosine similarity. This step is vital as we want to avoid heavy searching over all 2587 Misconceptions and top 25 retrieval can help us efficiently shrink the range and focus on more relevant misconceptions.

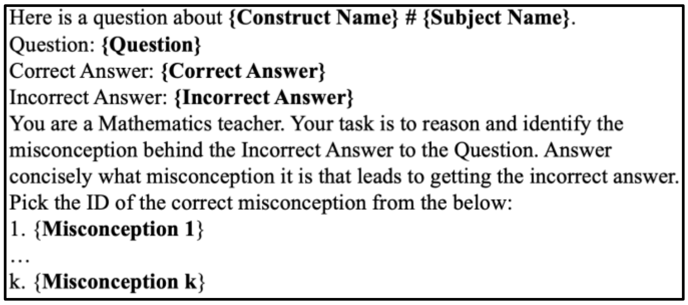
## Embedding Tuning (optional)

As the pre-trained LLM is not specialized in mathematical reasoning, fine tuning the embedding over the training dataset would be helpful. The most straightforward way would be supervised fine tuning LLM with techniques like LORA. However, limited by our computing resources, we try to fit a single-layer scoring model ***S*** that directly maps embedding into a matching score ***s = S(D,M)*** instead to finetune the final embedding. We expect ***S*** to output 1 when Distractors ***D***, Misconceptions ***M*** are matched, otherwise 0.

To train ***S***, we first examine the ranking of ground truth Misconception with vanilla embedding-based retrieval. The pre-trained Mistral model is already capable of matching correct Misconceptions for most Distractors, which means we need to focus on the "**hard**" Distractors that Mistral fails to identify the correct Misconceptions. To achieve this, we sample positive and negative ***(D,M)*** pairs for Distractions from their top25 relevant Misconceptions. For each Distractor ***D***, we sample 1 positive pair (***D,M*)*** where ***M**** is the Ground truth Misconception. Then for a hard Distractor, we sample 3 negative pairs (***D,M')*** where M' is randomly picked from the rest top 25 relevant Misconceptions, otherwise we only sample 1 negative pair. Finally, we sample 14073 pairs from the 3930 training Distractors. We train ***S*** with Adam optimizer, learning rate 1e-3 and 5 epochs.

We then evaluate Mistral Embedding and our scoring model ***S*** over the remaining 440 validation Distractors by retrieving top 25 Misconceptions with cosine similarity/output score, the result is included in the final table. Even though the scoring method ***S*** fails to outperform direct embedding retrieval due to its limited reasoning capability, we expect the training technique may also be helpful in future finetuning over LLM.

## Post-retrieval Misconception Selection



Here is a question about **{Construct Name} # {Subject Name}**.
Question: **{Question}**
Correct Answer: **{Correct Answer}**
Incorrect Answer: **{Incorrect Answer}**
You are a Mathematics teacher. Your task is to reason and identify the misconception behind the Incorrect Answer to the Question. Answer concisely what misconception it is that leads to getting the incorrect answer.
Pick the ID of the correct misconception from the below:
1. {**Misconception 1**}
…
k. {**Misconception k**}

After retrieving top k relevant misconceptions, we utilize LLM to do in-context reasoning, where Distractor and candidate Misconceptions are included in a single prompt and the LLM is required to select the most relevant one, to better exploit their capability. Here is the concrete prompt. In practice, we use Qwen-Instruct-32B to select the most relevant Misconceptions from the top 25 retrieval results. And this achieves about 0.09 leaderboard improvement compared to vanilla embedding retrieval.

## Result and Conclusion

| | **Validation Set** | **Public LeaderBoard** |
|---|---|---|
| Pretrained Mistral-7B | 0.842 | 0.353 |
| Pretrained Mistral-7B + Embedding Tuning | 0.571 | 0.200 |
| Pretrained Mistral-7B + Post-retrieval Selection | NaN | **0.447** |

The concrete empirical evaluation results are included in the table above. In this project, we created a distractor labeling system capable of identifying misconceptions behind the incorrect answers to the question by utilizing the reasoning capability of state-of-art open source LLMs. And design a process to mine hard training samples and reconstruct the training dataset, such techniques might be helpful for future LLM finetuning.