1. Problem Statement

In educational settings, students often harbor certain misconceptions that lead them to choose incorrect answers and hinder the learning process. Identifying these underlying misconceptions is crucial for teachers, as it allows for more targeted instruction and intervention, ultimately improving educational outcomes. To this end, Diagnostic Questions have been introduced as an effective tool. A Diagnostic Question is a multiple-choice question with four options: one correct answer and three distractors (incorrect answers) - each specifically crafted to capture a specific misconception.

However, manually tagging distractors with appropriate misconceptions is both time-consuming and prone to inconsistency among different human annotators. This inconsistency can undermine the accuracy and utility of the diagnostic process.

In this project, we propose to design an automatic 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.

2. Dataset

We use the Kaggle dataset Eedi to develop our model. It is composed of csv files for train/test dataset and a misconception mapping file including misconception id and descriptions. Following is a demonstration of train/test dataset format:

- QuestionId - Unique question identifier (int).
- ConstructId - Unique construct identifier (int).
- ConstructName - Most granular level of knowledge related to question (str).
- CorrectAnswer - A, B, C or D (char).
- SubjectId - Unique subject identifier (int).
- SubjectName - More general context than the construct (str).
- QuestionText - Question text extracted from the question image using human-in-the-loop OCR (str).
- Answer[A/B/C/D] Text - Answer option A text extracted from the question image using human-in-the-loop OCR (str).
- Misconception[A/B/C/D] Id - Unique misconception identifier (int). Ground truth labels in train.csv; we need to predict labels for test.csv.

3. Method

3.1 Objective

This project aims to employ existing open-source lightweight pre-trained LLMs (e.g. phi-3.5-instruct, bge-small-en-v1.5) for the prediction of misconception id given question content and chosen distraction. These models are expected to possess strong reasoning capability suitable for this task, given their training on internet-scale texts.

3.2 Evaluation Metric

To ensure that our system adequately captures the logic relationship between question, distraction, and misconception, we refer to the official evaluation metric of Eedi – Mean Average Precision@25 to evaluate the performance of our model:

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

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.