Name: **Mohammed Hany & Ahmed Samir** **MisrBase** **NTI**

Abstract

MisrBase is an AI-powered diagnostic tool designed to analyze student explanations in mathematics and detect conceptual misconceptions. Using a combination of Natural Language Processing (NLP) and machine learning (ML), the system classifies student responses across three tasks: correctness, category, and potential misconceptions. This report outlines the technical approach, model performance, challenges faced, and its alignment with Egypt Vision 2030 for educational reform and digital transformation.

1 Introduction

Identifying student misconceptions is critical to improving math education. Traditional assessments often focus on correct answers without uncovering why a student made an error. MisrBase addresses this gap by analyzing the language in open-ended student responses to detect common conceptual misunderstandings.

2 Objectives

MisrBase contributes to five national objectives, aligned with Egypt Vision 2030:

- **Revolutionize Curriculum Quality (Vision Pillar: Quality Education):** Provide the Ministry of Education with data-driven evidence on persistent national & regional misconceptions, enabling precise, evidence-based curriculum refinements for global competitiveness.
- **Empower the Egyptian Teacher (Vision Pillar: Building Digital Capacity):** Equip every Egyptian teacher with real-time, personalized diagnostic tools to identify why students struggle, saving time and enabling targeted interventions.
- **Build Foundational & Future Skills (Vision Pillars: Innovation & Technical Skills):** Move beyond rote memorization. Foster deep conceptual understanding, critical thinking, and articulate problem-solving – essential skills for Egypt's future workforce
- **Promote Equity & Inclusion (Vision Pillar: Equal Opportunities):** Identify and address learning gaps early, regardless of school location (urban/rural) or socioeconomic background, ensuring all Egyptian students build a strong math foundation.
- **Establish Egypt as a Regional EdTech Leader (Vision Pillar: Knowledge Economy):** Position MisrBase as a world-class, locally-developed AI solution for educational diagnostics, showcasing Egyptian innovation.

3 Alignment with Egypt Vision 2030

This work directly contributes to the goals of Egypt Vision 2030 in the areas of:

Table 1: Alignment with Egypt Vision 2030 Pillars

Pillar	**Contribution**
Economic Competitiveness	Builds a stronger STEM foundation, crucial for future innovation and high-tech industries. Data-driven curriculum ensures graduates meet global standards.
Knowledge, Innovation & Scientific Research	Uses cutting-edge AI/NLP developed in Egypt for Egypt. Creates valuable educational research datasets. Fosters an innovative EdTech ecosystem.
Social Justice & Inclusion	Democratizes access to high-quality diagnostic tools for all schools and teachers, reducing educational disparities. Early intervention supports struggling learners.
Education & Training	Directly enhances teaching quality and curriculum relevance. Shifts focus to deep understanding and critical thinking, moving beyond rote learning.
Digital Transformation	Embodies digital innovation in a core public service (education). Builds digital literacy for teachers and students through platform use.
Government Efficiency	Provides the MoE with unprecedented real-time data for evidence-based policy-making and resource allocation, optimizing educational investments.

4 Dataset Overview

The dataset originates from the Eedi platform, where students interact with Diagnostic Questions (DQs). These are multiple-choice questions (MCQs) specifically designed to identify common student misconceptions. Each DQ contains one correct answer and three distractors (plausible incorrect options).

After selecting an answer, students were sometimes prompted to provide a written explanation justifying their choice. These explanations are the central focus of the MAP (Misconception Annotation Project) dataset and serve as the primary input for model development.

Objective of the Dataset:
The goal is to train models that perform the following three tasks:

- Determine the correctness of the selected multiple-choice answer.
- Assess whether the explanation contains a mathematical misconception.
- Identify the specific misconception present, if any.

5 Data Structure

All question content, including mathematical expressions originally presented in image format, has been accurately extracted using a human-in-the-loop OCR process to ensure high fidelity.

Column Name	**Description**
`QuestionId`	Unique identifier for each diagnostic question.
`QuestionText`	The full text of the question presented to the student.
`MC_Answer`	The multiple-choice answer selected by the student.
`StudentExplanation`	The student's open-ended explanation justifying their answer choice.
`Category`	**[Train only]** Combined label indicating correctness and explanation quality (e.g., `True_Misconception`).
`Misconception`	**[Train only]** The specific misconception identified in the explanation. If not applicable, set to `NA`.

Table 2: Columns in the dataset

6 Feature Engineering and Target Preparation

This project employs a structured preprocessing pipeline to extract both linguistic and mathematical reasoning features from students' open-ended explanations. The following functions were implemented to prepare data for machine learning models:

6.1 Mathematical and Linguistic Feature Extraction

The function `extract_mathematical_features(text)` computes handcrafted features related to mathematical syntax, reasoning language, and uncertainty cues. These include:

- **Basic text features:** character length, word count, and sentence count.
- **Mathematical content:** presence of fractions, decimals, negatives, and operations ($+$, $-$, $\times$, $\div$, $=$).
- **Numerical patterns:** total number of numeric tokens present.
- **Reasoning indicators:** such as `because`, `so`, `therefore`, and `since`.
- **Uncertainty expressions:** such as `think`, `maybe`, `guess`, and `probably`.
- **Common misconception terms:** `bigger`, `smaller`, `more`, `less`.

Each of these features is encoded as binary or numeric and stored in a dictionary per student explanation.

6.2 Feature Preparation Pipeline

The function `prepare_features(df, tfidf_vectorizer, question_encoder, answer_encoder)` combines multiple feature types:

- **Mathematical features:** Extracted using the function above and stored in a dataframe.
- **TF-IDF textual features:** Applied to student explanations to capture lexical patterns. A sparse matrix is converted into dense feature vectors.
- **Categorical encodings:** Both question text and multiple-choice answers are label-encoded.

- **Question/Answer metadata:** Character lengths of questions and answers are included as numerical features.

All features are concatenated into a final dataframe used for model training.

6.3 Target Variable Preparation

The function `prepare_targets(train_df)` extracts three key targets from the training set:

1. **Correctness (binary):** Encoded from the `Category` column prefix (True = 1, False = 0).
2. **Category type (multi-class):** Extracted from the suffix in `Category` (e.g., Correct, Misconception, Neither), and label-encoded.
3. **Specific misconception (multi-class):** Derived from the `Misconception` column. NA values are handled as a separate class.

The label encoders for category type and misconceptions are saved and used consistently across training and inference.

6.4 Feature Summary

The resulting feature space includes:

- Handcrafted linguistic/mathematical features (e.g., 20+ binary/numeric columns)
- TF-IDF features (typically hundreds of dimensions)
- Encoded categorical features for questions and answers
- Length-based metadata

This hybrid feature representation supports both traditional ML models (e.g., LightGBM) and integrated approaches in downstream tasks.

7 Model Architecture

For this task, we fine-tuned the **roBERTa-base** model, a robustly optimized BERT variant trained on a larger corpus. Given its superior performance on downstream classification tasks, roBERTa-base was chosen as the encoder to process the student explanations and extract meaningful contextual embeddings.

The model performs multi-task classification to simultaneously predict:

1. The correctness of the student's selected answer.
2. Whether the explanation contains a misconception.
3. The type of misconception, if applicable.

The final architecture consists of the pre-trained roBERTa-base followed by:

- A dropout layer with probability 0.1 to prevent overfitting.
- A fully connected classification head with Softmax for each task.

8 Training Details

The model was trained using the HuggingFace `transformers` library with the `Trainer` API. The training configuration is as follows:

- **Model:** roBERTa-base
- **Epochs:** 5
- **Train Batch Size:** 16
- **Eval Batch Size:** 16
- **Learning Rate:** 1×10^{-5}
- **Warmup Steps:** 100
- **Weight Decay:** 0.01
- **Evaluation Strategy:** Every 100 steps
- **Logging Steps:** 25
- **Save Strategy:** Every 100 steps (best model saved)
- **Metric for Best Model:** `eval_f1`
- **Hardware:** Dual NVIDIA T4 GPUs on Kaggle platform
- **Precision:** FP16 enabled if GPU available

Early stopping and checkpoint saving were applied to ensure the best performing model was preserved.

9 Results

The model was evaluated using key performance metrics to assess its capability in classifying student answers based on correctness, conceptual category, and presence of misconceptions.

Table 3: Individual Model Performance

Metric	**F1 Score**
Correctness Classification	0.9995
Category Classification	0.8068
Misconception Detection	0.8895

Mean Average Precision at 3 (MAP@3)

To further evaluate the model's ranking capability—especially useful when recommending multiple relevant misconception labels—we computed the Mean Average Precision at 3 (MAP@3). It is defined as:

$$\text{MAP@3} = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{\min(n,3)} P(k) \times \text{rel}(k)$$

Where: - U is the total number of examples, - $P(k)$ is the precision at rank k, - $\text{rel}(k)$ is a binary indicator (1 if the prediction at rank k is relevant, else 0), - n is the number of predictions returned by the model.

The MAP@3 score for the model was **0.7902**, indicating that in the top 3 predictions, the model ranked relevant misconception labels with high accuracy. This is especially important in educational applications where multiple misconceptions may be possible, and recommending the most likely ones assists teachers in rapid intervention.

10 Conclusion

In this project, we developed and fine-tuned a multi-task learning model based on `roBERTa-base` to analyze student responses on the Eedi platform. The model accurately identifies correct answers, misconceptions, and specific misunderstanding types using natural language explanations provided by students.

Our approach not only achieves strong predictive performance but also opens the door for educational technology tools that adapt to learners' cognitive needs. This supports Egypt's Vision 2030 goals in education, innovation, and digital transformation. Future directions include integrating explanation generation and expanding the model to other STEM subjects.