An Intelligent System for Detecting Student Misconceptions in Mathematics

INTELLIGENT SYSTEM FOR DETECTING STUDENT MISCONCEPTIONS IN MATHEMATICS APPLICATION PAPER. JULY 12, 2025

Steven Navarro Parrales
20221020048

Juan David Amaya Patiño
20221020057

David Santiago Garcia Galeano
20231020158

***Abstract*—This paper presents the development of an intelligent system designed to automatically detect and classify student misconceptions in mathematics based on multiple-choice distractors. The system was developed as a response to the Kaggle competition "Eedi – Mining Misconceptions in Mathematics" and leverages Natural Language Processing (NLP), machine learning, and local Large Language Models (LLMs), specifically Ollama. Using a modular and feedback-oriented architecture, the system processes question-answer pairs, computes semantic similarity between distractors and misconceptions, and employs a validation loop to improve prediction accuracy. The system integrates soft systems thinking, chaos mitigation strategies, and cybernetic feedback for error correction. Results demonstrate high classification accuracy and strong MAP@25 performance metrics across varying dataset sizes, confirming the system's robustness, adaptability, and educational relevance. The successful implementation validates the viability of automated diagnostic tools in educational environments and provides a foundation for future personalized learning systems.**

***Index Terms*—misconception detection, natural language processing, large language models, educational technology, semantic similarity, machine learning**

I. Introduction

The identification and classification of student misconceptions in mathematics education represents a critical challenge in modern educational systems. Traditional approaches to identifying these systematic errors in student reasoning have relied heavily on manual analysis by educators, a process that is both time-consuming and prone to inconsistency. This paper presents an intelligent system designed to automate the detection and classification of student misconceptions using advanced Natural Language Processing (NLP) techniques and Large Language Models (LLMs).

The system was developed in response to the Kaggle competition "Eedi – Mining Misconceptions in Mathematics," which challenges participants to predict which misconceptions are associated with specific distractors in mathematics questions. These distractors often reveal common misunderstandings that students have about fundamental mathematical concepts, making their accurate identification crucial for developing effective educational interventions.

Our approach addresses the core problem of scaling educational feedback and correction mechanisms in a way that is both efficient and personalized. The system leverages the semantic understanding capabilities of LLMs to capture nuanced meanings in natural language and generalize across varying question structures and error types. By implementing a modular architecture that incorporates semantic filtering, machine learning classification, and LLM-based validation, the system demonstrates the potential for automated diagnostic tools to transform educational assessment practices.

The integration of systems theory principles, including cybernetic feedback loops and chaos mitigation strategies, ensures the system's stability and adaptability under varying input conditions. The successful implementation of this system marks a significant advancement in the application of artificial intelligence to educational diagnostics, offering educators and researchers a powerful tool for understanding and addressing student misconceptions at scale.

II. Materials and Methods

The development of the intelligent misconception detection system followed a systematic approach grounded in systems engineering principles and advanced machine learning methodologies. The system architecture was designed to handle the complexity and variability inherent in educational data while maintaining robust performance across different dataset configurations.

The initial phase focused on comprehensive system analysis and design, where project requirements were carefully outlined based on the Kaggle competition specifications. The system was conceptualized as a modular architecture comprising interconnected components: data preprocessing, semantic similarity analysis, classification, and validation. This modular approach enabled independent development and testing of each component while ensuring seamless integration.

The core methodology integrated multiple technological approaches. Python 3.10 was selected as the primary programming language, providing robust capabilities for data processing and machine learning implementation. The system employed SentenceTransformers for generating semantic embeddings, enabling accurate computation of similarity between distractor texts and misconception descriptions. A RandomForest classifier was implemented using scikit-learn to perform multi-label classification based on semantic similarity scores and linguistic features.

The integration of Ollama as a local Large Language Model represented a key innovation in the system design. Unlike cloud-based LLMs, Ollama enabled local deployment, ensuring data privacy, cost control, and consistent performance. The LLM was integrated at two critical stages: post-filtering tagging to associate distractors with relevant misconceptions, and validation to verify and refine predictions through a feedback loop.

Data preprocessing involved comprehensive text cleaning, tokenization, and filtering to ensure data quality and system stability. Semantic similarity computation used cosine similarity between sentence embeddings, with a threshold of 0.15 applied to filter low-affinity pairs. The system was designed to handle datasets of varying complexity, with testing conducted across configurations of 100, 500, and 1000 misconceptions to evaluate scalability and robustness.

The evaluation framework centered on the Mean Average Precision at 25 (MAP@25) metric, as specified by the competition requirements. This metric measures the quality of ranked predictions by considering both the relevance and order of the top 25 predicted misconceptions. Additional internal metrics included classification accuracy, processing time, and system stability measures to ensure comprehensive performance evaluation.

III. Results and Experiments

The implementation and evaluation of the intelligent misconception detection system yielded comprehensive results that demonstrate its effectiveness across multiple performance dimensions. The system successfully processed question-answer pairs and generated ranked lists of misconceptions with high accuracy and consistency.

The primary evaluation metric, MAP@25, achieved a best score of 0.811 on the simulated test set containing 1000 misconceptions. This performance level indicates strong predictive capability in identifying and ranking relevant misconceptions for given distractors. The system maintained consistent performance across different dataset sizes, with standard deviation of MAP@25 scores remaining within ± 0.021 across simulation rounds, demonstrating robust stability.

Classification accuracy using the RandomForest model ranged from 82% to 100% across different test configurations, with the semantic similarity threshold of 0.15 proving optimal for filtering low-relevance pairs. The system successfully identified 73% of filtered distractors as high-affinity pairs, indicating effective semantic matching between distractors and misconceptions.

The integration of Ollama as the validation component showed promising results, with only a 6% rejection rate during the validation phase. This low rejection rate, combined with an average LLM inference time of 1.2 seconds per prompt, demonstrates the efficiency of the local LLM approach. The system processed 100 distractors in approximately 2.5 minutes, indicating suitable performance for practical educational applications.

System behavior analysis revealed adaptive labeling capabilities, where the LLM dynamically updated or corrected misconception labels during the feedback loop. When trained on reduced datasets, the model exhibited conservative labeling behavior, prioritizing fewer but higher-confidence predictions. The cybernetic feedback structure effectively mitigated error propagation by reprocessing low-confidence predictions through the LLM tagger.

The system consistently detected key mathematical misconceptions, including misunderstandings about triangle angle sums, confusion between area and perimeter concepts, incorrect square root operations, and inverse operation errors. These recurring detections indicate that the system captures meaningful cognitive errors rather than superficial text matches.

Performance metrics across different dataset complexities showed zero-crash runtime stability, confirming system resilience under varied input conditions. The modular design enabled parallel processing and independent component testing, supporting the system's scalability for larger educational deployments.

IV. Conclusion

The development of this intelligent system for detecting student misconceptions in mathematics represents a significant advancement in the application of artificial intelligence to educational diagnostics. The system successfully addresses the challenge of automating misconception identification through a sophisticated integration of Natural Language Processing, machine learning, and Large Language Model technologies.

The modular architecture, grounded in systems theory principles, demonstrates the effectiveness of combining semantic filtering with LLM-based reasoning to achieve accurate and scalable misconception detection. The use of Ollama as a local LLM provides distinct advantages in terms of data privacy, cost control, and deployment flexibility while maintaining competitive performance levels.

Results confirm the system's technical viability and educational relevance, with high MAP@25 scores and robust classification accuracy across varying dataset complexities. The system's ability to consistently identify meaningful mathematical misconceptions validates its potential for integration into adaptive learning platforms and intelligent tutoring systems.

The successful implementation of cybernetic feedback loops and chaos mitigation strategies ensures system stability under realistic data perturbations, addressing a critical concern in educational AI applications. The modular design supports future enhancements and adaptations, making the system suitable for expansion to other mathematical domains or educational subjects.

While certain limitations remain, including domain-specific focus and the need for real-time optimization, the system provides a solid foundation for automated educational diagnostics. Future work should focus on multilingual support, cross-domain generalization, and classroom deployment interfaces to maximize the system's practical impact.

This project contributes to both the academic understanding of systems analysis in educational technology and the practical development of AI-powered diagnostic tools. The successful integration of advanced NLP techniques with educational assessment represents a promising direction for enhancing personalized learning at scale.

References

[1] Kaggle Competition: Eedi – Mining Misconceptions in Mathematics. Available: https://www.kaggle.com/competitions/eedi-mining-misconceptions-in-mathematics/

[2] Scikit-learn: Random Forests. Available: https://scikit-learn.org/stable/modules/ensemble.html#random-forests

[3] Sentence-BERT: Semantic Textual Similarity. Available: https://www.sbert.net/docs/usage/semantic_textual_similarity.html

[4] Ollama: Local Large Language Models. Available: https://ollama.ai/

[5] Pandas Documentation. Available: https://pandas.pydata.org/docs/

[6] NumPy Documentation. Available: https://numpy.org/doc/

[7] Python unittest Documentation. Available: https://docs.python.org/3/library/unittest.html