

Automated Approaches for Scoring Math Misunderstandings in Student Self-Explanations

Scott A. Crossley¹, Bethany Rittle-Johnson¹, Rebecca Adler¹, L Burleigh², Kelley Durkin¹, Jules King², Meg Benner²

¹Vanderbilt University

²The Learning Agency

{scott.crossley, bethany.rittle-johnson, rebecca.adler, kelley.durkin}@vanderbilt.edu
{lburleigh, jules, meg}@the-learning-agency.com

Abstract

This study reports on an open data science competition based on a benchmark dataset of mathematics misunderstandings, comprising over 52,000 mathematics explanations written to justify answer choices from 15 multiple-choice questions that were labeled by expert human annotators. Competitors were tasked with correctly classifying the explanations and any misunderstandings. By combining stable validation methods with efficient inference and enriched training data, top teams achieved high accuracy scores that correctly classified the explanations as correct, a misunderstanding, or neither and, if it did have a misunderstanding, what type of misunderstanding it was.

Student proficiency in mathematics has declined in the past decade, and pandemic-related school disruptions have left lasting impacts on the nation and world's education systems. There is strong enthusiasm about the potential of Artificial Intelligence (AI) to help address such educational challenges. However, to harness the promise of AI for improving mathematics education, AI models need to be able to not only diagnose if a student answered a question correctly but also understand potential mathematics misunderstandings in the student's thinking. This is because effective mathematics instruction does not just teach correct concepts and procedures, but also directly addresses mathematics misunderstandings, including misconceptions. Misunderstandings interfere with students' ability to learn correct

concepts and procedures and can persist for many years (Byrd et al., 2015). Directly addressing and countering students' misunderstandings improves learning outcomes, including when instruction occurs on digital learning platforms (Barbieri et al., 2019; Durkin & Rittle-Johnson, 2012, 2015; Huang et al., 2008).

However, to our knowledge, no work has focused on developing large language models (LLMs) to score student self-explanations for math misunderstandings. Corollaries have been the use of LLMs for scoring the accuracy of open-ended student responses, such as automated essay scoring (Ramesh & Sanampudi, 2022) and short answer scoring (Bonthu et al., 2021) including open-ended mathematics responses (Liu et al., 2023; Morris et al., 2025).

To address this, the current study reports on the results of a data science competition to detect mathematics misunderstandings using the Misunderstandings in Math Explanations (MIME) dataset introduced in Rittle-Johnson et al. (2025). The goal of the competition was to solicit innovative AI modeling approaches that can detect patterns of misunderstandings in student mathematics responses. Participants were required to classify 1) whether a selected multiple-choice answer to a math question was correct, 2) whether a self-explanation of that multiple-choice answer was correct, a target misunderstanding, or neither (e.g., vague or irrelevant), and 3) if the self-explanation contained a misunderstanding, what type of misunderstanding was present. The last step, which included models classifying self-explanation into predefined misunderstanding

categories is important because it can help provide insights into student thinking. Models from the competition were evaluated on their accuracy in detecting the steps of the process.

1 Methods

1.1 MIME Dataset

The competition was based on the MIME dataset. The dataset includes an initial dataset of over 38,000 student explanations written over 15 mathematics questions covering key middle-grade mathematics topics (grades 4-8) from Eedi, a mathematics learning platform based in the U.K. All questions and multiple-choice selections were written by mathematics teachers. Each item begins with a multiple-choice question with 1 correct and 3 incorrect options, with incorrect options meant to capture known misunderstandings as much as possible. Selected items covered core topics in middle-grade mathematics curricula, including topics in rational numbers and in equations and functions. Included items also were required to have at least 1,000 explanation responses available and each response had to contain at least 5 English words (with one exception).

An additional 14,368 synthetic explanations were generated using Coedit-XL to supplement the 38,095 original explanations. A maximum of one augmented explanation was generated by paraphrasing an authentic student explanation. To ensure authenticity and approximate observed features of authentic student writing, one spelling error was randomly injected into 50% of the augmented explanations, and 50% of the augmented explanations were fully lowercased.

Explanations (both synthetic and original) were labeled by human annotators using standardized annotation rubrics and procedures developed by experts. For each explanation in the dataset, annotators determined if the explanation was correct, contained a misunderstanding, or was neither (e.g., vague or irrelevant explanations, including ones that just repeated the multiple-choice answer without elaboration). If the explanation included a misunderstanding, the annotators determined what type of misunderstanding was present. Neither was a catch-all label for any explanation that did not meet the criteria for other labels. The final MIME dataset comprises 52,463 explanations and metadata in tabular format.

Further details about the MIME dataset including how the misunderstanding categories were developed, how they were used in scoring, whether they were item-specific or shared across items, division of training, validation, and test sets, and annotator inter-rater reliability are reported in Rittle-Johnson et al. (2025). The dataset is available at https://github.com/scrosseye/mime_dataset.

1.2 Competition Tasks and Objectives

The data science competition using the MIME dataset was held on Kaggle, which hosts data science and machine learning challenges. The competition files and winning models are posted on <https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/leaderboard>. The competition was open to data scientists, researchers, and educators interested in applying computational methods to real-world educational data. Participants ranged from academic teams to professionals who use machine learning in industrial application (e.g., software engineers, data scientists, AI engineers, etc.). The participants brought diverse expertise in predictive modeling, natural language processing, and educational technology to the competition. The competition lasted three months and attracted over 1,850 teams, 2,175 entrants, and resulted in over 39,760 entries. The developers of the six top-performing models were given cash prizes ranging from \$20,000 USD for 1st place to \$5,000 for 4th-6th place (Total prize amount available was \$55,000).

Participants were tasked with predicting the type of understanding present in each student's answer. This required classification of whether the multiple-choice answer selected was correct and whether the explanation was correct, a target misunderstanding, or neither. The objective was to develop models capable of identifying correct and different types of incorrect reasoning to support automated feedback and educational interventions. Participants were given ~37,000 responses for training models (i.e., 70% of the data was in the training set). The responses were from all 15 items. The training data file included: (a) QuestionId - Unique question identifier, (b) QuestionText - The text of the question, (c) MC_Answer - The multiple-choice answer the student selected, (d) StudentExplanation - A student's explanation for choosing a specific multiple-choice answer, (e)

Category - A classification of the relationship between a student's multiple-choice answer and their explanation (e.g., `True_Misunderstanding`, which indicates a correct multiple-choice answer selection accompanied by an explanation that reveals a misunderstanding), and (f) `Misunderstanding` - the mathematical misunderstanding identified in the student's explanation for answers when Category contains a misunderstanding. There were 6 possible Categories created by crossing True (correct multiple-choice answer) or False (incorrect multiple-choice answer) with three explanation labels: Correct, Misunderstanding, or Neither.

For the overall task, participants were asked to develop a model that performed 3 steps:

- Determines whether the selected multiple-choice answer was correct. (True or False)
- Assesses whether the explanation was correct, contained a misunderstanding, or was neither (i.e., Correct, Misunderstanding, or Neither).
- Identifies the specific misunderstanding present when applicable.

A validation set with 15% of the data (~7,800 responses not included in the training set and drawn from all 15 items) was used to provide competitors with their models' real-time score. A test set with 15% of the data (~7,800 responses drawn from all 15 items) was used to evaluate the final competitors' model performance.

Mean Average Precision @ 3 (MAP@3) was used as a scoring metric. In MAP@3, the models were given three chances to predict the ground truth (human) annotation for a given student response. Models received diminishing scores after each incorrect prediction. For example, if the human labels B and the model predicts [B, A, C] in a sequence, it scores 1.0. If it predicts [A, B, C], it scores 0.5. If it predicts [A, C, B], it scores 0.33. A high MAP@3 score means that the model consistently placed the correct label within the top three predictions, most often as the primary choice.

1.3 Baseline Model

A simple baseline model used for comparison purposes was created using a DeBERTa-v3-large encoder model with hyperparameters and training arguments (learning rate = 1e-05, batch size=8, weight decay = 0.03) with cross entropy as the loss function for four epochs. On the held-out test set, the model reported a weighted MAP@3 = 0.80.

2 Results

Top submissions achieved high accuracy in classifying misunderstandings, as well as correct explanations. The top models achieved Mean Average Precision (MAP@3) scores slightly above 0.948, well above the competition's benchmark of 0.80, reflecting the strength of data science approaches to applying AI modeling to mathematics education. Additionally, the top 1,000 competitor models were able to achieve at least a MAP@3 scores of 0.94, which increases the likelihood that a variety of LLMs' can perform well on this task. Winning models are posted online on Kaggle and are freely available for use as baseline scoring models in research or in learning platforms.

Subsequent interviews with winning teams and reviews of discussion forums and code on Kaggle indicated a few core challenges in the competition that lowered performance. These included noisy and ambiguous labels, data imbalance across misunderstanding types, and computational constraints. Noisy and ambiguous labels arose because some explanations were difficult to classify (even for humans), especially the ambiguous and catch-all label *neither*. Indeed, *neither* was the most difficult label for models to predict. Data imbalance arose because the frequency of particular misunderstandings ranged from 0.2% to 35% of explanations. Lastly, complex LLMs, such as a 72 billion parameter LLM used in some submissions, incurred substantial computational costs leading many winning teams to use smaller, open-weight reasoning models, with QWEN and Z.ai models being common.

Successful approaches used by competitors generally shared three key characteristics. First, successful approaches used robust validation strategies such as ensembling, k-fold cross validation, and multi-seed validation for handling noisy and ambiguous labels. For example, the first-place winner noted that multi-seed validation was "the most critical aspect of this competition" for handling noisy labels. Second, teams used advanced inference engineering, including quantization which reduces model weights from 16-bit to 4-bit. This reduced the size of a 72B model from 144GB to roughly 35–40GB.

Teams also used multi-stage inference to deploy multiple higher parameter models (32-72B) within the competition's time constraints. For example, the third-place winner ran expensive 72B models only on low-confidence predictions. Successful

approaches also used data augmentation to create synthetic data to expand the set of misunderstanding labels to boost performance. For example, the second-place team generated 80K synthetic explanations using commercial LLMs and expanded brief misunderstanding labels into detailed descriptions.

Three innovative approaches emerged from the three top-performing teams, who all leveraged ensembles of LLMs. The first-place team framed the task as a ‘suffix classification task’ where each label candidate was encoded as an input token, and, for each prediction, all possible label candidates were concatenated as a string. Then, custom FlexAttention masks were applied so that each suffix candidate (all the possible answers) only attended to its shared prefix, defined as the context which includes the question text, the multiple-choice options, the correct answer, misunderstanding labels, the student answer, and the student explanation. The last-token features of each list of suffixes and prefixes was used to get logits for cross-entropy classification. This approach framed the problem as asking the LLM to choose from a fixed list of possibilities (the label candidates). So, instead of asking the language model to generate an answer, it transformed the problem into a multiple-choice question format within the model’s architecture.

The second-place team generated additional training data by averaging predictions from multiple trained models (called ‘soft labels’), then trained with combined loss on both to potentially account for label ambiguity. The soft labels were probability scores given to each possible category from an ensemble of predictions from multiple, already-trained models. The LLM used the hard label to learn the ground truth and the soft labels to learn the nuances (e.g., which inputs are difficult, where the labels are ambiguous, and which misunderstandings are more plausible).

The third-place team added auxiliary tasks. For instance, they required training models to simultaneously predict related sub-problems like whether the multiple-choice answer was correct and what type of reasoning error occurred in the explanation, providing additional learning signals that improved the main classification task. This allowed for a single system that became smarter across all tasks by learning from the shared relationships between the tasks. By learning from

interrelated tasks, the model also avoided overfitting (memorizing specific answers).

3 Discussion and Conclusion

This study introduces a data science competition run using the MIME dataset. This competition yielded thousands of models that could accurately classify a range of conceptual and procedural misunderstandings, as well as correct explanations. The competition results demonstrate the potential for data-driven approaches to empower teachers and digital learning platforms to identify and address gaps in student understanding using efficient LLM pipelines.

3.1 Research Implications

The data science competition yielded key insights for developing AI models that can accurately classify student math misunderstandings. Successful teams combined stable validation methods with efficient inference and enriched training data to generate models that could accurately classify students’ thinking on the first attempt in the vast majority of cases. Even greater model accuracy was achieved using three particularly innovative approaches (e.g., a suffix classification task with custom FlexAttention masks).

One key limitation was that solutions were often optimized for competition-specific features. In particular, the dataset was based on multiple-choice questions with options aligned with many of the misunderstandings. Data on whether students answered multiple-choice responses correctly were available in the competition, which likely made detecting potential misunderstandings easier. Further, competition teams only needed to predict the subset of misunderstandings previously observed for a question. An open inquiry is how well the top models from the competition are transferable to new mathematics content, including different mathematics questions, other potential misunderstandings, and open-ended items without related multiple-choice questions. For this last concern, because the MIME dataset is openly available, new models can be trained that do not include multiple-choice responses.

The state-of-the-art large language models developed based on the MIME dataset can support sophisticated digital learning platforms to help detect mathematics misunderstandings in students’ explanations. The competition’s outcomes point

toward a future in which adaptive and automated feedback systems can become integral to mathematics instruction. The use of automated feedback systems that incorporate student misunderstandings would help teachers maximize their time and personalize learning by targeting instructional approaches based on students' specific knowledge without requiring excessive additional effort.

Overall, the results from the data science competition provide a foundation for building future AI scoring models aligned with mathematics standards and assessments centered on student misunderstandings. Misunderstandings are considerably more difficult for LLMs to identify than correct understanding (Liu et al., 2023). As models trained on misunderstanding data become more sophisticated, they will contribute to improved diagnostic tools for teachers and more precise feedback for students, advancing both research and classroom practice. This includes supporting more precise knowledge tracing in intelligent tutoring systems so that LLMs can steer instruction in the most effective way after students make errors (Liu et al., 2022; Scarlatos et al., 2025). Such models also need to be evaluated for their accuracy in the classroom, the instructional usefulness of feedback, teacher interpretation of feedback, classroom implementation, and effects on student learning. A similar approach has been very successful in the domain of writing, supporting automatic generation of feedback and dialogic chatbots for essay and short-answer writing (Bonthu et al., 2021; Crossley et al., 2024, 2025; Ramesh & Sanampudi, 2022).

3.2 Future Research

Future efforts are needed to explore the practical deployment and generalization capacities of high-performing models from the data science competition. The top 1,000 models on the private leaderboard all performed above 0.941, which was only 0.007 points below the winning model. This suggests the potential to find a 'balanced' model that offers a good tradeoff between accuracy, speed, and reliability. Such a model would be best suited for real-world use, where the algorithm must run efficiently and consistently identify misunderstandings. The competition established a baseline performance metric which can be used as a reasonable benchmark for what constitutes strong

model accuracy while remaining practical in terms of computational requirements and inference time.

Future research should also assess the generalization capability of these models beyond the competition's scope. Testing top-performing models on newly annotated mathematics questions using the same annotation scheme would reveal which approaches can generalize to questions they were not trained on. This is particularly relevant given that the training, validation, and test sets all shared the same 15 multiple-choice questions, which may have allowed for competition-specific optimization that may not transfer to broader applications. Understanding which modeling strategies maintain robust performance across different question sets and response formats would be valuable for deploying systems for detecting misunderstandings in educational settings.

Acknowledgments

Vanderbilt University and The Learning Agency would like to thank the Gates Foundation, Schmidt Futures, and the Walton Family Foundation for their support in making this work possible. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Gates Foundation, Schmidt Futures, and the Walton Family Foundation.

References

- Barbieri, C. A., Miller-Cotto, D., & Booth, J. L. 2019. Lessening the Load of Misconceptions: Design-Based Principles for Algebra Learning. *Journal of the Learning Sciences*, 28(3), 381–417. <https://doi.org/10.1080/10508406.2019.1573428>
- Bonthu, S., Sree, S. R., & Prasad, M. H. M. K. 2021. Automated Short Answer Grading Using Deep Learning: A Survey. In *Machine Learning and Knowledge Extraction International Cross-Domain Conference, CD-MAKE* (pp. 61–78). Springer-Verlag. https://doi.org/10.1007/978-3-030-84060-0_5
- Byrd, C. E., McNeil, N. M., Chesney, D. L., & Matthews, P. G. 2015. A specific misconception of the equal sign acts as a barrier to children's learning of early algebra. *Learning and Individual Differences*, 38, 61–67. <http://dx.doi.org/10.1016/j.lindif.2015.01.001>
- Crossley, S. A., Baffour, P., Burleigh, L., & King, J. 2025. A large-scale corpus for assessing source-based writing quality: ASAP 2.0. *Assessing Writing*,

- 65, 100954.
<https://doi.org/10.1016/j.asw.2025.100954>
- Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Benner, M., & Boser, U. 2024. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61, 100865.
<https://doi.org/10.1016/j.asw.2024.100865>
- Durkin, K., & Rittle-Johnson, B. 2012. The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, 22, 206–214.
<https://doi.org/10.1016/j.learninstruc.2011.11.001>
- Durkin, K., & Rittle-Johnson, B. 2015. Diagnosing misconceptions: Revealing changing decimal fraction knowledge. *Learning and Instruction*, 37, 21–29.
<https://doi.org/10.1016/j.learninstruc.2014.08.003>
- Huang, T.-H., Liu, Y.-C., & Shiu, C.-Y. 2008. Construction of an online learning system for decimal numbers through the use of cognitive conflict strategy. *Computers & Education*, 50, 61–76.
- Liu, N., Sonkar, S., Wang, Z., Woodhead, S., & Baraniuk, R. G. 2023. Novice Learner and Expert Tutor: Evaluating Math Reasoning Abilities of Large Language Models with Misconceptions.
<https://arxiv.org/abs/2310.02439>
- Morris, W., Holmes, L., Choi, J. S., & Crossley, S. 2025. Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models. *International Journal of Artificial Intelligence in Education*, 35(2), 559–586.
<https://doi.org/10.1007/s40593-024-00418-w>
- Ramesh, D., & Sanampudi, S. K. 2022. An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Rittle-Johnson, B., Adler, R., Durkin, K., Burleigh, L., King, J., & Crossley, S. A. (2025). Detecting Math Misconceptions: An AI Benchmark Dataset. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con)* pp. 20–24, National Council on Measurement in Education (NCME).
- Scarlato, A., Fernandez, N., Ormerod, C., Lottridge, S., & Lan, A. 2025. SMART: Simulated Students Aligned with Item Response Theory for Question Difficulty Prediction. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 25071–25094). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2025.emnlp-main.1274>