Detecting Math Misconceptions: An AI Benchmark Dataset

Bethany Rittle-Johnson[1], Rebecca Adler[1], Kelley Durkin[1], L Burleigh[2], Jules King[2], Scott Crossley[1],

[1]Vanderbilt University, [2]The Learning Agency,

Correspondence: bethany.rittle-johnson@vanderbilt.edu

Abstract

To harness the promise of AI for improving math education, AI models need to be able to diagnose math misconceptions. We created an AI benchmark dataset on math misconceptions and other instructionally relevant errors, comprising over 52,000 explanations written over 15 math questions that were scored by expert human raters. A data science competition based on the dataset will lead to state-of-the-art large language models for detecting math misconceptions.

1 Background

Student proficiency in math has declined in the past decade, and pandemic-related school disruptions have left lasting impacts on the nation's and world's education systems (National Center for Education Statistics (NCES), 2022). To harness the promise of AI for improving math education, AI models need to be able to diagnose students' potential math misconceptions and other instructionally relevant errors. Misconceptions are "any student conception that produces a systematic pattern of errors" (Smith III et al., 1994). They often form as people attempt to "assimilate…new information into their existing conceptual structures" (Stafylidou and Vosniadou, 2004). For example, students often inappropriately generalize concepts and procedures learned about whole numbers to fractions and decimals (i.e., whole number bias; Durkin and Rittle-Johnson, 2012, 2015; Stafylidou and Vosniadou, 2004). Misconceptions interfere with students' ability to learn correct concepts and procedures and can persist for many years (e.g., Byrd et al., 2015). Other instructionally relevant errors include attending to irrelevant information and incomplete solution procedures.

Directly addressing and countering students' misconceptions improves learning outcomes, including on digital learning platforms (Barbieri et al., 2019; Durkin and Rittle-Johnson, 2012; Huang et al., 2008). However, more research and development infrastructure is needed to ensure that teachers and digital learning platforms can integrate information about math misconceptions into product development, research, and instruction. We hosted a workshop that gathered mathematics cognition researchers and digital learning platform developers together to align needs and priorities, which helped guide our work.

One needed tool is benchmark datasets of math misconceptions and other instructionally relevant math errors. This would enable assessment of how well AI models perform, providing an objective way to compare different AI models and platforms, ensuring transparency, accountability, and suitability for use in education (Hodeem, 2024).

1.1 Aims

This research methods paper introduces an AI benchmark dataset on math misconceptions and other instructionally relevant errors. The dataset is the focus of the MAP - Charting Student Math Misunderstandings data science competition hosted on Kaggle, ending in October 2025, to generate large language models that can detect math misconceptions.

2 Sample

The dataset comprises over 52,000 student explanations written over 15 math questions covering key middle-school math topics. We used a dataset from Eedi, a math learning platform based in the U.K., which has been used by over 250,000 teachers in 16,000 schools. All questions and feedback messages were written by math teachers. Each item begins with a multiple-choice question with 1 correct and 3 incorrect options, with incorrect options meant to capture known misconceptions and other instructionally relevant errors as much as

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 20–24

possible. We sampled from instances that included a follow-up prompt for an open-ended explanation of why students selected their answer. We selected items that covered core topics in the U.S. middle-school math curriculum, primarily topics in rational numbers and in equations and functions, and for which there were at least 1000 explanation responses available with at least 5 English words (with one exception). Explanations were screened for personally identifiable information before being shared. No demographic information about the participating students was available.

To ensure a meaningful sample of each misconception code and increased explanations for extensive rater training and norming, 14,368 synthetic explanations were generated using Coedit-XL to supplement the 38,095 original explanations (27% synthetic explanations). A maximum of 1 augmented explanation was generated by paraphrasing an authentic student explanation. Coedit-XL tends to provide the correct spelling, punctuation, and capitalization in paraphrasing students' explanations. To better reflect the writing characteristics of authentic student submissions, one spelling error was randomly injected into 50% of the augmented explanations, and 50% of the augmented explanations were fully lowercased. To confirm the realism of synthetic explanations, an expert reviewed a sample of explanations that were partially real and partially synthetic without an indication of the source. Coedit-XL parameters were iterated on until the expert was unable to differentiate the sample.

3 Methods

Students' explanations (both synthetic and original) were scored by human raters using standardized scoring rubrics and procedures. A scoring rubric for each item was developed by three experts in math cognition and misconceptions, drawing on past research on misconceptions as much as possible. The rubric identified criteria for correct explanations and 2-4 potential instructionally-relevant errors, as outlined in Table 1. Each explanation could only receive one code. Raters went through extensive norming prior to independent rating. Raters were primarily undergraduate or graduate students with prior experience teaching or tutoring children in mathematics.

4 Results

The final dataset comprises 52,463 explanations and metadata in tabular format. The dataset contains student ID numbers, item IDs, explanations, and human-assigned codes. Twenty-four percent of explanations were coded by two raters, with high inter-rater reliability (Cohen's Kappa .70-.90). One item with weak inter-rater reliability was dropped. Inter-reliability was also assessed separately for real and synthetic explanations, which resulted in similar Cohen's Kappa values for each item with large enough synthetic explanation sample sizes for reliable statistical evaluation (real vs synthetic Cohen's Kappa differences: 0.005 - 0.15).

Across the 15 items, 27% of explanations had evidence of a potential misconception, and the frequency of particular misconceptions ranged from 0.2% to 35% of explanations. For example, when asked to calculate $\frac{2}{3} \times 5$, 23% of explanations indicated a misconception that the whole number is converted to a fraction with both its numerator and denominator as that whole number. 41% of explanations were correct (range: 18-74%).

This dataset supports the development and evaluation of state-of-the-art large language models that can detect potential misconceptions, including the MAP - Charting Student Math Misunderstandings data science competition hosted on Kaggle, ending in October 2025. Submitted models will be evaluated with the Mean Average Precision @ 3 metric. Winning models will be posted on Kaggle and can be used as baseline scoring models. One potential limitation is that the models may be overly tuned to particular concepts that may be over-represented because synthetic explanations were paraphrased from authentic explanations. To get a better understanding of the models, prediction accuracies can be evaluated for authentic and synthetic explanations, as well as for each misconception code.

Item Topic	Annotated sample size	Error 1	Error 2	Error 3
Fraction Representation	6,963	Believes numerator and denominator of a fraction indicate two separate numbers [WNB]	Incomplete steps: fails to simplify fraction [Incomplete]*	
Adding Fractions	3,994	Adding numerators and denominators without finding common denominator [Adding across]	Finds common denominator and adds numerators [Denominator-only change]*	Creates equivalent fractions, and adds numerator and denominator [Incorrect Equivalent fraction addition]
Finding Fraction of Set #1	4,023	Attending to irrelevant feature [Irrelevant]	Incomplete steps: Calculates unit fraction only [Incomplete]	Calculates fraction for the wrong target [Wrong fraction]*
Finding Fraction of Set #2	2,206	–	Incomplete steps: Calculates unit fraction only [Incomplete]	Calculates fraction for the wrong target [Wrong fraction]*
Equivalent fractions	5,204	Attending to irrelevant feature [Irrelevant]	Treating fraction as 2 separate numbers [WNB]	Additive thinking: finds difference [Additive]*
Dividing fractions	4,476	Multiplies to divide by whole number [Mult.]*	Swaps the divisor and dividend [SwapDividend]	Flips the dividend (not divisor) and keeps it as division [FlipChange]
Multiplying two Fractions	2,528	Division instead of multiplication [Division]	Subtracts the provided number [Subtracts]*	
Fraction whole number multiplication	4,411	Multiplies numerator and denominator by the whole number (instead of just the numerator) [Duplication]*	Inverts the whole number multiplier [Inversion]	Adds rather than multiplies [Wrong operation]
Solve for y	3,080	Treats y as a missing digit, rather than a variable [Not variable]*	Transforming problem to addition problem [Adding terms]	Applies wrong operation (i.e., multiplies rather than divides) [Inverse]
Decimal Magnitude	3,320	Believes whole numbers are larger than numbers with decimals [Whole numbers larger]*	Believes longer numbers are bigger [Longer is bigger]	Believes zeroes do not add magnitude information [Ignores zeroes][a]

Continued on next page

Item Topic	Annotated sample size	Error 1	Error 2	Error 3
Polygon Sides	1,695	Believes there is not enough information to solve problem [Unknowable]*	Does not use correct formula, and instead divides the total interior angle sum by one interior angle [Interior]	Believes a polygon is defined by having a certain number of sides (5 or 6) [Definition]
Subtracting a negative	4,365	Ignores negative signs and adds them back at the end [Tacking]	Incorrect application of two negatives makes a positive [Two negatives is positive]*	
Functional thinking	3,727	Uses the first term of the output as the coefficient of the rule [firstterm]	Calculates the n+1 term, rather than the n+2 term [wrong term]*	
Proportions	968	Reverses proportional reasoning by multiplying instead of dividing [Multiplying by 4]*	Incorrect base rate [Base rate]	
Odds	1,503	Does not understand the range of probability is 0 to 1 [Scale]*	Believes events with probability $\neq 1$ are certain [Certainty]	

Table 1. Error Categories and Frequencies in the Math Misconceptions AI Benchmark Dataset by Item. Notes: *Most frequent error type for each item; [a]Fourth code was: Believes fewer digits after the decimal point, the larger the number is [Shorter is bigger].

5 Conclusion

To harness the promise of AI for improving math education, AI models need to be able to diagnose students' potential math misconceptions and other instructionally relevant errors. We have created an AI benchmark dataset on math misconceptions covering a variety of middle-school math topics that will be publicly available, along with baseline scoring models. Although this dataset is based on explanations primarily from students in the U.K., their explanations align with misconceptions and correct ways of thinking identified in the research literature conducted primarily in the U.S. and Canada. State-of-the-art large language models based on this dataset will support digital learning platforms' ability to detect math misconceptions, and multiple digital learning platforms are interested in adding this capability. Detecting misconceptions is necessary for them to be addressed, and directly addressing and countering students' misconceptions improves learning outcomes (Barbieri et al., 2019; Durkin and Rittle-Johnson, 2012; Huang et al., 2008).

References

Christina A. Barbieri, Dana Miller-Cotto, and Julie L. Booth. 2019. Lessening the load of misconceptions: Design-based principles for algebra learning. *Journal of the Learning Sciences*, 28(3):381–417.

Caroline E. Byrd, Nicole M. McNeil, Dana L. Chesney, and Percival G. Matthews. 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.

Kelley Durkin and Bethany Rittle-Johnson. 2012. The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, 22(3):206–214.

Kelley Durkin and Bethany Rittle-Johnson. 2015. Diagnosing misconceptions: Revealing changing decimal fraction knowledge. *Learning and Instruction*, 37:21–29.

Hodeem. 2024. Ai: An overview of common llm benchmarks. https://dev.to/hmcodes/ai-an-overview-of-common-llm-benchmarks-3i7b.

Tzu-Hua Huang, Yuan-Chen Liu, and Chia-Ya Shiu. 2008. Construction of an online learning system for decimal numbers through the use of cognitive conflict strategy. *Computers Education*, 50(1):61–76.

National Center for Education Statistics (NCES). 2022. NAEP long-term trend assessment results: Reading and mathematics—Age 9. https://www.nationsreportcard.gov/ltt/?age=9.

John P. Smith III, Andrea A. diSessa, and Jeremy Rochelle. 1994. Misconceptions preconceived: a constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, 3(2):115–163.

Stamatia Stafylidou and Stella Vosniadou. 2004. The development of students' understanding of the numerical value of fractions. *Learning and Instruction*, 14(5):503–518.