JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 12(1), 50–64. https://doi.org/10.18608/jla.2025.8621

Learning to Love LLMs for Answer Interpretation: Chain-of-Thought Prompting and the AMMORE Dataset

Owen Henkel[1], Hannah Horne-Robinson[2], Maria Dyshel[3], Greg Thompson[4], Ralph Abboud[5], Nabil Al Nahin Ch[6], Baptiste Moreau-Pernet[7] and Kirk Vanacore[8]

Abstract
This paper introduces AMMORE, a new dataset of 53,000 math open-response question-answer pairs from Rori, a mathematics learning platform used by middle and high school students in several African countries. Using this dataset, we conducted two experiments to evaluate the use of large language models (LLM) for grading particularly challenging student answers. In experiment 1, we use a variety of LLM-driven approaches, including zero-shot, few-shot, and chain-of-thought prompting, to grade the 1% of student answers that a rule-based classifier fails to grade accurately. We find that the best-performing approach — chain-of-thought prompting — accurately scored 97% of these edge cases, effectively boosting the overall accuracy of the grading from 96% to 97%. In experiment 2, we aim to better understand the consequential validity of the improved grading accuracy by passing grades generated by the best-performing LLM-based approach to a Bayesian Knowledge Tracing (BKT) model, which estimated student mastery of specific lessons. We find that modest improvements in model accuracy can lead to significant changes in mastery estimation. Where the rule-based classifier misclassified the mastery status of 6.9% of students across completed lessons, using the LLM chain-of-thought approach reduced this to 2.6%. These findings suggest that LLMs could be valuable for grading fill-in questions in mathematics education, potentially enabling wider adoption of open-response questions in learning systems.

Notes for Practice

- The AMMORE dataset is a new resource for learning analytics research into student math practice drawn from an understudied, real-world educational context.
- We find that when trying to grade the most ambiguous student answers, using a large language model–based approach can slightly outperform sophisticated text-processing–based approaches.
- We also find that modest improvements in grading accuracy at the question-level leads to substantially improved estimates of student mastery across lessons.

Keywords: Large language models (LLMs), formative assessment, math education
Submitted: 09/08/2024 — **Accepted:** 13/01/2025 — **Published:** 23/03/2025

Corresponding author [1]Email: owen.henkel@education.ox.ac.uk Address: Department of Education, University of Oxford,15 Norham Gardens, Oxford, United Kingdom. ORCID iD: https://orcid.org/0009-0001-8850-067X
[2]Email: hannah.horne-robinson@risingacademies.com Address: Rising Academies Ghana, Egli Altezza Plaza, New Bortianor, Accra, Ghana.
[3]Email: maria@tangibleai.com Address: TangibleAI, San Diego, CA, USA.
[4]Email: greg@tangibleai.com Address: TangibleAI, San Diego, CA, USA.
[5]Email: rabboud@levimath.org Address: Learning Engineering Virtual Institute, Digital Harbor Foundation, 1045 Light St, Baltimore, MD, 21230, USA. ORCID iD: https://orcid.org/0000-0002-2332-0504
[6]Email: ch.nabil.nahin@gmail.com Address: University of Minnesota, Department of Educational Psychology, 250 Education Sciences Building, 56 East River Road, Minneapolis, MN, 55455, USA. ORCID iD: https://orcid.org/0000-0002-0202-1724
[7]Email: baptiste@levi.digitalharbor.org Address: Digital Harbor Foundation, 1045 Light St, Baltimore, MD, 21230, USA. ORCID iD: https://orcid.org/0009-0006-9424-455X
[8]Email: kirk.vanacore@gmail.com Address: University of Pennsylvania Graduate School of Education, Philadelphia, Pennsylvania, USA. ORCID iD: https://orcid.org/0000-0003-0673-5721

1. Introduction

Effective learning systems rely on accurate, real-time evaluation of student knowledge states to optimize learning trajectories and provide targeted support (Black & Wiliam, 2010; Gikandi et al., 2011). While multiple-choice questions enable rapid knowledge state assessment within digital platforms, they can inadvertently measure test-navigation skills rather than true

knowledge states and may not capture the full complexity of student understanding (Johnson & Green, 2006; Rupp et al., 2006). Open-response questions, particularly in mathematics, offer richer insights into student thinking and problem-solving processes, providing learning systems with more nuanced data about knowledge construction and application (Magliano & Graesser, 2012; O'Neil & Brown, 1998). Most common in mathematics is fill-in responses, where grading is simple for exact matches but becomes more challenging when flexibility is required. The complexity of automatically evaluating these responses presents significant technical challenges for learning platforms, particularly when students express correct solutions through diverse mathematical notations or explanations, or when chat-based interfaces lead students to respond using natural language, which can be difficult to interpret. Hence, supporting automated, accurate assessment of student answers is a major goal of such platforms because it allows students to receive immediate feedback and creates better learning experiences (Black & Wiliam, 2010; Funk & Dickson, 2011).

The challenge of automatically marking open-response fill-in math questions presents a critical opportunity for learning analytics (LA) to support the evaluation of student performance during learning, which is key to many essential functions of educational technology (e.g., mastery evaluation, automated feedback, reporting performance to educators). While many student responses can be evaluated through straightforward text processing techniques, a considerable subset of responses pose complex challenges — they may be formatted unconventionally, contain extraneous information, or require nuanced interpretation (Botelho et al., 2023; Hahn et al., 2021). The prevalence of these challenging responses, particularly in large-scale online learning platforms, creates a pressing need for more sophisticated assessment approaches.

While various automated approaches have been attempted (Allen et al., 2014; Burrows et al., 2015; Crossley et al., 2019), most have required extensive technical expertise and large datasets (Mayfield & Black, 2020; Pulman & Sukkarieh, 2005). Recent evidence suggests that LLMs can accurately evaluate responses with minimal prompt engineering (Gilardi et al., 2023; Henkel et al., 2024), potentially enabling more frequent and effective formative assessment while reducing educator workload. However, questions remain about their reliability across diverse educational contexts and their ability to handle increasingly complex assessment scenarios. Furthermore, the field lacks sufficient publicly available educational datasets to thoroughly evaluate these approaches. This paper makes two contributions in response to these gaps.

First, we introduce a novel dataset, the African Middle-School Math Open REsponse (AMMORE) dataset, which consists of 53,000 responses to middle school math questions from students in West Africa. The data for AMMORE was collected from Rori, an AI-powered chat-based math tutor that allows students in West Africa to independently practise math concepts free of charge. The dataset's rich structure, which includes question-level data, user IDs, learning standard designators, and students' self-reported age, enables various potential analyses, such as investigating student skill mastery across lessons, analyzing the relative difficulty of specific questions or lessons across students, or exploring how the judgments of different grading models compared to those of humans. Beyond automated scoring research, this dataset offers LA researchers a unique opportunity to study student engagement patterns, learning trajectories, and interaction behaviours within a free digital math education platform. The temporal and sequential nature of the data allows for investigation of how students navigate content, persist through challenges, and utilize automated feedback — critical questions for understanding and improving online learning environments. As noted by Motz et al. (2023), there is a surprising lack of research in LA focused on learning outcome data derived from applied learning systems. AMMORE provides a unique opportunity to explore learning outcomes from diverse students in a real-world educational context. Second, LA research has historically focused primarily on data from the North America and Europe (Cechinel et al., 2020), and this data set can facilitate LA research in a less commonly studied region (West Africa) and context, informal chat-based learning.

Second, we conducted an extensive empirical evaluation of LLM-based approaches to grade a difficult-to-grade subset of AMMORE. We explore various automated methods — including string matching, text processing, and different LLM prompting techniques — to evaluate their accuracy and consistency in assessing student responses. We find that LLM-based approaches, particularly chain-of-thought prompting (CoT), outperform traditional methods in grading accuracy, demonstrating their ability to handle the complexity and variability of student responses to fill-in math questions. The superior performance of LLM-based methods is especially evident in cases where students provide correct answers in unexpected formats or use equivalent mathematical expressions. We also explore whether relatively modest improvements in answer scoring accuracy at the individual question level can lead to significantly more accurate estimates of student concept mastery and more optimal next lesson recommendations. Our findings also suggest that the use of LLM-based grading could encourage wider adoption of open-response questions in digital learning platforms, leveraging their pedagogical benefits without increasing the grading burden on educators.

2. Description of Learning Platform

Rising Academies (2024), an educational network based in Ghana, has created Rori, an AI-powered chat-based math tutor available on WhatsApp. Students chat with Rori using natural language like any WhatsApp contact and are expected to write their responses to math questions using the mobile keyboard. Rori's curriculum is built upon the comprehensive, evidence-

ISSN 1929-7750 (online).

based Global Proficiency Framework (GPF; USAID, 2019). The GPF was developed by key global education organizations to create uniform global standards for reading and mathematics across the world. The GPF covers grades 1 to 9, aligns with national standards globally, and is linked across grade levels. The GPF organizes content by domain constructs, then subconstructs, and then into specific skills that a student in each grade should be able to demonstrate. For example, the domain "Numbers and Operations" has a topic "Integers and Exponents" with skills such as "Add and subtract" and "Multiply and divide."

The Rori curriculum has one or more lessons for each skill in the math GPF, with over 700 lessons to date. Each lesson includes a brief student-friendly explanation of the skill and 10 practice questions of the same difficulty. Many of these questions require fill-in responses, a decision made for pedagogical reasons. If students answer a question incorrectly, they are first shown a hint to help them solve the question and are given another chance to solve it, as shown in Figure 1. If their second attempt is unsuccessful, they are shown a worked solution and then given the next question. When students finish a lesson, they are encouraged to continue with the next as lain out by the GPF, meaning they incrementally increase in difficulty. Rori will suggest students move either backwards or forwards in the curriculum if they find a lesson too difficult or too easy.[1]

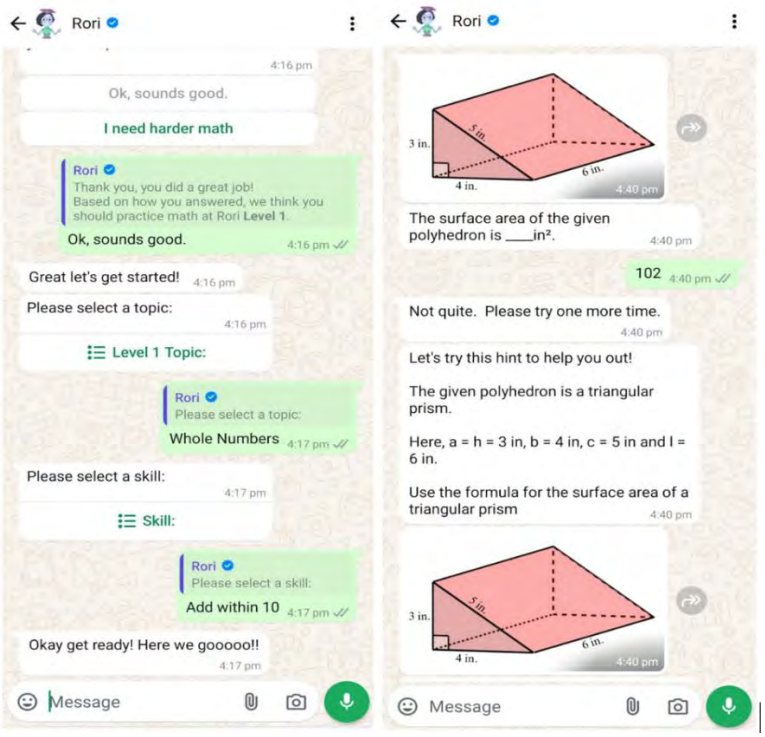

Figure 1. Example of lesson experience.

3. Prior Work

3.1. Challenges of Evaluating Student Work in Online Learning Platforms

Accurately estimating a student's current knowledge state and tracking their progress and understanding of the subject enables learning systems to deliver a better learning experience (Abdelrahman et al., 2023; Chrysafiadi & Virvou, 2013). For example, student modelling can be used for making key decisions such as which problem a student should attempt, how much practice is needed to master a skill before moving to a more advanced topic, and when to provide immediate feedback to struggling students (Cukurova et al., 2022; Feng et al., 2009). In the context of a learning environment, even a small number of mis-graded answers can lead to vastly different judgments of student ability when aggregated across questions (Dey et al., 2024; Pelánek, 2015). In the specific case of Rori, there are diverse fill-in answer types, including fractions, floating-point numbers, and expressions with exponents. Because the interface is conversational, students also frequently type messages that are not attempted answers or mix a natural language response and their answer. A core challenge of this type of learning system is being able to distinguish between student messages that are and are not answer attempts, and between correct and incorrect answer attempts. Table 1 presents a range of student responses for a given question.

[1] For more context, see Henkel (2024).

Table 1. Example Student Answers and Labels

question_id: G7.N2.2.3.6

question_text: Fill in the missing number: 1/5 × 2/3 = _ /15

expected_answer: 2

student_id	student_answer	model_grade	human_grade
514	2	correct	correct
1073	Hold am solving it	other	wrong
876	is 2	correct	correct
1203	30	wrong	wrong
324	2/15	wrong	correct

While Rori was already using a relatively sophisticated mix of text processing and NLP classification models to interpret and score student responses, which had already undergone several rounds of improvement, after human review we found that approximately 1% (1,186) of classifications were false negatives and found no false positives. From a pedagogical perspective, it is important to avoid misclassifying correct student responses (i.e., false negatives) as much as possible — particularly in independent learning environments — as telling a student they made a mistake when they were in fact correct can lead to confusion and frustration (Hsu et al., 2021; Rajendran et al., 2019). For example, from the responses in Table 1, requiring a perfect answer match would mean only student 514's answer would be accepted; being too permissive and only students 1073 and 1203 would be marked incorrect. Looking at the response given by student 324, an expert human reviewer can understand that the student did the core mathematical operation correctly and gave the full answer rather than only giving the missing number. While there might be a pedagogical reason to encourage the student to use the correct formatting, treating their answer as wrong would be suboptimal. This long tail of particularly subtle and difficult-to-grade student responses may benefit from combining more traditional NLP-based approaches with LLMs.

3.2. Potential of Generative LLMs for ASAG

The current generation of LLMs, including ChatGPT, GPT-4, Claude, Llama, Mistral, and Gemini, underwent various "instruction fine-tuning" steps to enhance their usability and ability to generalize to new tasks, often with minimal exposure to examples (Ouyang et al., 2022). This also improved their interpretation of human-written natural language instructions (i.e., prompting), allowing non-technical users to make requests and adapt a model to new tasks by modifying their prompts, rather than requiring further training or fine-tuning (Stiennon et al., 2022). Current LLMs can perform various linguistic tasks that previously required the use of task-specific, fine-tuned LLMs (Kojima et al., 2023; Wei et al., 2022). Therefore, it is unsurprising that evidence is growing that LLMs can be used for certain types of grading tasks (Kortemeyer, 2023).

There is a growing body of research on using generative LLMs to evaluate student work. Morjaria et al. (2024) found that ChatGPT graded six short answer assessments from an undergraduate medical program similarly to a single expert rater. Cochran et al. (2022) found that GPT-4 successfully graded student answers to high school science questions. However, Kortemeyer (2023) found that LLMs fell short in certain aspects of grading introductory physics assignments. Recently, LLMs have been used to autograde student responses to math questions. Botelho et al. (2023) used a pre-trained Sentence-BERT to assess student responses to open-ended math questions, predict student scores, and recommend appropriate feedback. Shen et al. (2023) created MathBERT, which outperformed prior methods and BERT on various math tasks including grading responses to fill-in math questions. Injeti et al. (2024) used MathBert to specifically grade algebraic questions. A review by Schneider et al. (2024) concluded that "while 'out-of-the-box' LLMs provide a valuable tool to offer a complementary perspective, their readiness for independent automated grading remains a work in progress." More exploration is needed into the ability of generative LLMs to grade responses to math questions, and there is a growing collection of publicly available datasets that could be used for this purpose.

3.3. Overview of Existing Short Answer Datasets

While there are several math question datasets in the literature (see Table 2 for a more detailed overview), they present some limitations that undermine their relevance for understanding student behaviour on real-world learning platforms. Some datasets (e.g., MATH) contain questions and correct answers but do not contain information about how students answered. Other datasets (e.g., EEDI and MathE) contain information about student multiple-choice responses. Of the datasets listed below, only ASSISTments contains information allowing researchers to track progression through a curriculum. None of these datasets contain information from students from lower-resourced and underrepresented populations. These limitations are the main motivation behind our dataset AMMORE, which we discuss in more detail in Section 4.

Table 2. Summary of Publicly Available Math Datasets

Dataset	Topic	Answers / Age	Country	Response Type	Number of Responses
MATH	Competition Mathematics	No / N/A	N/A	Open Response	12,500
GSM8K	Primary School Mathematics	No / N/A	N/A	Open Response	8,000+
MathE	College Mathematics	Yes / No	Multiple	Multiple Choice	9,546
COMAT	Primary, Middle & High School Mathematics	Yes / No	United States	Conversations & Open Response	188
EEDI	Primary, Middle & High School Mathematics	Yes / No	United Kingdom	Multiple Choice	17 million+
NAEP	Grade 4, Grade 8 Mathematics	Yes / Yes	United States	Constructed Response	250,000+
ASSISTments	Middle School Mathematics	Yes/ No	United States	Multiple Choice & Open Response	1,000,000+

4. The AMMORE Dataset

In this section, we present the African Middle-School Math Open REsponse (AMMORE) Dataset, which contains 53,298 student answers to open-response practice questions, assembled from a subset of math practice sessions on Rori from 2,508 at-home users that took place between 1 January and 30 April 2024.

4.1. Description of Dataset

The AMMORE dataset is composed of student responses to math questions from Rori lessons from grade levels 6 to 9 in the domains of "Algebra" and "Number and Operations." Within these domains, there are 151 unique lessons — corresponding to a specific knowledge component and/or skill — which cover 35 distinct constructs (please see repository for a more complete description and a data dictionary). Each response in our dataset was scored by a pre-existing, rules-based classification model, or answer evaluation API, native to Rori, which classifies answer attempts as "correct," "wrong," or "other." The latter was typically returned when a student entered something besides an answer attempt, such as a voice note or a sticker. Humans then manually reviewed these classifications. Two math educators inspected the student responses classified as "wrong" or "other" to determine where they disagreed with the model. Any ambiguous responses were discussed until agreement was reached. This means that the dataset also has a ground truth score for each student response. The human raters also noted any errors in the Rori questions or expected responses.

Table 3 shows the structure of the dataset. Each student response is paired with the corresponding question, the expected response, a ground-truth correct/incorrect score, the specific learning standard evaluated by the question, the time the student answered, and a UID number that can be used to link student responses across the dataset.

Table 3. Structure of AMMORE dataset

Summary Information		Example attributes of single entry	
Total Answers	53,031	lesson	G9.N5.2.1.1
Correct Answers	34,668	question_number	2
Incorrect Answers	15,278	question_text	3^2 + 3^1 = __
Other Answers	3,085	expected_answer	12
Unique Students	2,508	student_response	=6+6 =12
Grade Levels Covered	6–9	model_grade	wrong
Domains Covered	Algebra, Numbers and Operations	human_grade	correct
Number of Lessons	151	time	1/9/24 7:57
Learning Constructs	35	user_id	17

The dataset also includes matched but anonymized demographic data on the 2,508 users, such as when they first started using Rori, their country code, self-reported age, number of messages they sent, and active days on Rori. At-home users tend to come from Nigeria, Ghana, and South Africa and are mostly between the ages of 10 and 30 and could be using their own or a family member's phone to access Rori.

4.2. Potential Uses of the AMMORE Rising Dataset

The dataset's structure enables various potential analyses. For example, 1) investigating student mastery across lessons, 2) analyzing the relative difficulty of lessons for students, or 3) exploring how the classification model's judgments compared to those of human raters.

For example, as discussed above, a lesson is a set of 10 interchangeable questions of equivalent difficulty level focusing on a specific learning standard from the GPF. If we were to posit that a student could be considered to have mastered the skill associated with a lesson if they get 80% of the questions correct, we would find that students "mastered" 48% of lessons in the dataset. To further this analysis, one could consider student progress at the construct level, which is a collection of closely related lessons. The dataset includes 151 different lessons covering 35 constructs. Also, because the same student practises skills at different grade levels, it is possible to compare student age to the grade level of the lesson they are practising to estimate whether students are performing at grade level. Our dataset's lessons span grades 6 to 9. Yet another approach would be to use this dataset to test different analytics approaches, such as Bayesian Knowledge Tracing (BKT), which we explore in experiment 2, or other mastery prediction models. The rich data available, including question-level responses and progression through lessons over time, makes this dataset particularly suitable for such analyses.

These are just a few potential uses for this novel dataset. The combination of detailed student responses, demographic information, and curriculum structure provides a unique opportunity for researchers to explore various aspects of learning analytics, from individual student progress to broader trends in mathematical skill development across grade levels.

5. Experiment 1: LLM-Based Approaches to Grading Math Questions

Using a carefully curated subset of difficult-to-grade student responses from the AMMORE dataset, we investigate six different automatic grading strategies, ranging from simple string matching to sophisticated LLM-based methods, evaluating their respective performance relative to human scores. We also consider how consistent the models are between repeated runs, if the prompting strategy affects the intra-rater reliability between the model's responses, and how prompting strategy impacts the model response time. Our analysis aims to shed light on the potential of these approaches to improve grading accuracy, particularly when dealing with diverse answer types and formatting variations.

5.1. Experimental Design

From the larger AMMORE dataset, we create a smaller dataset, which we refer to as AMMORE-hard. This dataset is composed of difficult-to-grade student responses, which we used to evaluate the performance of different automatic grading strategies. The resulting dataset comprises 4,463 responses, including 1,528 unique non-trivially correct answers and 2,935 unique, non-trivially wrong answers.

AMMORE-hard was created using the following steps: 1) remove answers labelled as "other" by a human labeller; 2) remove duplicate occurrences where question, expected answer, and student answer were identical, leaving only one occurrence of each unique combination; 3) remove trivially correct answers, where the student response was identical to the expected answer; 4) remove trivially wrong answers, where the expected answer was one character long and the student's response was one character long (mostly multiple-choice questions with wrong answer); 5) remove wrong answers, where the student response was an integer but was different from the integer expected. The result is a subset of student answers that require some nontrivial amount of interpretation.

Then using the six different approaches described in Table 4, we scored student responses from AMMORE. To make a prediction, each approach was given the same information from the dataset: the question text, the expected answer to the question, and the student's response. The evaluation approach would predict if the answer were "correct" or "wrong." The resulting prediction was recorded. At the time of writing, the model with the strongest performance score on math benchmarks is OpenAI's GPT-4 (Chatbot Arena, 2024). Hence, each experiment of a prompt approach used GPT-4 as the LLM. Its temperature was set to 0 to reduce the variability of model outputs. No student demographic information was fed to the LLM, nor was it shown the human labels of a student answer.

Table 4. Different Approaches to Grading Student Answers

"Naive" string matching	Simple rule-based evaluation of matching the expected answer with the student response.
Text processing	Evaluation with additional text substitutions and symbolic evaluations.
LLM Zero-shot prompting	Evaluation using an LLM prompt without specific examples.
LLM Few-shot prompting	Evaluation using an LLM prompt with a small set of examples.
LLM Chain-of-thought prompting	Evaluation using an LLM prompt instructing it to show its reasoning process.
Naive string matching, text processing, and zero-shot prompting	Evaluation proceeded through the evaluations until a correct answer was found or all three evaluations had run: simple rule-based evaluation, text substitutions and symbolic evaluations, and an LLM prompt without examples.

5.2. Prompting Strategy

We employ a relatively simple prompting strategy, as the task is straightforward. The base part of the prompt was similar across all strategies. The zero-shot prompt included a description of the core task and slots for the dataset values. The few-shot prompt added three examples of correct answers. These examples represented common student response patterns of equivalent answers: 1) where a student wrote the answer and 2) where a student wrote out their work to arrive at the answer. Instead of providing examples, the chain-of-thought prompt instructed the model to think step-by-step and present a rationale for the classification chosen. The chain-of-thought evaluation used the DSPy framework (2023), which dynamically created a chain-of-thought prompt. Table 5 shows the prompts for each strategy.

Table 5. System Prompts Used in Experiment

Zero-shot Prompt	Few-shot Prompt	Chain-of-thought Prompt
You are a math assistant. You are evaluating whether a student's submission to a math question is right or wrong. The student may have submitted a correct answer in a variety of acceptable, equivalent ways. You must tell whether their submission correctly solves the problem or whether their submission contains a valid answer that is equivalent to the expected answer. If the student's submission is correct or equivalent, write, "yes." If the submission is incorrect and not equivalent, write, "no." You should only write "yes" or "no." ## Question {question} ## Expected Answer {expected_answer} ## Student Submission {student_message}	You are a math assistant. You are evaluating whether a student's submission to a math question is right or wrong. The student may have submitted a correct answer in a variety of acceptable, equivalent ways. You must tell whether their submission correctly solves the problem or whether their submission contains a valid answer that is equivalent to the expected answer. If the student's submission is correct or equivalent, write, "yes." If the submission is incorrect and not equivalent, write, "no." You should only write "yes" or "no." ## Examples ### Example 1: The student gave their work and showed the correct answer. - Question: Solve for z in the proportion: 9/3 = 27/z. - Expected Answer: 9 - Student Submission: 9/3=27/a.9×z=3×27.9z/9=91/9.z=9 - is_correct: yes ### Example 2: The student wrote the correct answer option and its value. - Question: 9 / ___ = 0.25 A) 18 B) 36 C) 81 D) 72 - Expected Answer: B - Student Submission: B.36 - is_correct: yes ## Question {question} ## Expected Answer {expected_answer} ## Student Submission {student_message}	You are a math assistant. You are evaluating whether a student's submission to a math question is right or wrong. The student may have submitted a correct answer in a variety of acceptable, equivalent ways. You must tell whether their submission correctly solves the problem or whether their submission contains a valid answer that is equivalent to the expected answer. Use the following format. Question: the math question Expected Answer: the student's response to the question Reasoning: Let's think step-by-step in order to produce the correct answer We... Answer: correct_answer if the student correctly solves the problem or whether their submission contains a valid answer that is equivalent to the expected answer, wrong_answer otherwise Question: {question} Expected Answer: {expected_answer} Student Answer: {student_answer} Reasoning: Let's think step-by-step in order to solve the equation {question}

To establish a baseline and evaluate the individual prompt strategies, we first implemented a simple text processing pipeline using regular expressions. This pipeline included normalizing case, removing extra whitespace, standardizing number formats (e.g., converting written numbers like "ninety-nine" to "99"), and applying basic string-matching rules. For each student response, we first processed both the expected answer and student answer through this pipeline, then compared them for exact matches. Responses that matched exactly after processing were labelled as correct, while non-matches were labelled as incorrect. These automated labels were then compared against expert human-annotated ground truth labels to evaluate the baseline performance. For the prompt strategy evaluations, we used this same dataset but instead passed the original question text, expected answer, and student answer through OpenAI's API using our various prompting approaches. The script recorded all evaluation outputs (i.e., the predicted class) for each method.

5.3. Results

Table 6 shows the results of the six approaches. As mentioned earlier, each answer evaluation would label a student's response as "correct" or "wrong." These predictions were compared against the label assigned by a human rater. In Table 6, a result closer to one indicates that the human label and the prediction were similar (e.g., both labelled a student answer as "wrong_answer"). A lower score would indicate that the human label and the predicted label differed (e.g., the human label marked "correct_answer" and the predicted label "wrong_answer").

In Table 6, we report a set of widely used metrics in classification problems that measure model performance after accounting for imbalanced classes in the dataset: precision, recall, and F1 score (Banerjee et al., 1999). Precision measures the proportion of correctly identified positive cases among all predicted positives, recall indicates the proportion of actual positive cases correctly identified, and F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's accuracy. We calculate these metrics separately for both correct and incorrect student answers to assess model performance across both response categories. We also report the Kappa scores, which are chance-adjusted metrics of agreement, with values ranging from -1 to 1. A value of 1 indicates perfect agreement, 0 suggests that the agreement is only what would be expected by chance, and a value of less than 0 indicates agreement worse than random chance. While there are several different measures of chance-adjusted agreement, because we are evaluating 2-class ratings (wrong/correct), we use Cohen's Kappa.

Table 6. Performance of Answer Evaluation Approaches on 2-Class Task

	Prediction	Accuracy	Precision	Recall	F1	Kappa
String Matching	Wrong	0.79	0.76	0.99	0.86	0.44
	Correct	0.79	0.97	0.39	0.56	
Text processing	Wrong	0.96	0.96	0.97	0.97	0.90
	Correct	0.96	0.94	0.93	0.94	
LLM Zero-shot	Wrong	0.94	0.93	0.98	0.95	0.86
	Correct	0.94	0.96	0.85	0.90	
LLM Few-shot	Wrong	0.93	0.91	0.99	0.95	0.83
	Correct	0.93	0.97	0.81	0.88	
LLM Chain-of-thought	Wrong	0.97	0.97	0.98	0.98	0.93
	Correct	0.97	0.96	0.94	0.95	
Text processing and LLM Zero-shot	Wrong	0.96	0.97	0.97	0.97	0.92
	Correct	0.96	0.94	0.95	0.95	

Our analysis revealed that all evaluation techniques generally performed better when attempting to predict a wrong answer rather than a correct answer. The lower recall and F1 scores observed in the zero-shot and few-shot results suggest that these techniques have difficulty handling the variety of unique correct answer formats present in the dataset. This limitation led to more false negatives, which is particularly problematic in educational scenarios where failing to recognize correct answers can impact student assessment.

Surprisingly, enhancing the prompt with examples did not necessarily lead to improved performance. When evaluating student messages, we found that combining text processing with LLM zero-shot prompting achieved similar performance to the Chain-of-Thought (CoT) prompting approach. The strong performance of CoT prompting suggests an improved capability of the latest LLMs to deal with mathematical tasks and solve math problems, as demonstrated in the example presented in Table 7.

Table 7. Output of Different LLM (GPT-4o) Strategies on a Sample Question-Answer Pair

question_id:	G6.N1.2.2.1.7
question_text:	3^2 + 3^1 = __
expected_answer:	12
student_response:	=6+6 =12
human_category:	correct_answer
llm_cot_category:	wrong_answer
llm_cot_rationale:	We need to evaluate the student's response to see if it correctly solves the problem or contains a valid answer that is equivalent to the expected answer. First, let's solve the problem step-by-step: - Calculate \(3^2\): \(3^2 = 9\) - Calculate \(3^1\): \(3^1 = 3\) - Add the results: \(9 + 3 = 12\) The expected answer is 12. Now, let's examine the student's answer: - The student wrote, "=6+6 =12" The student's answer is not a correct step-by-step solution to the problem. The student incorrectly broke down the calculation into \(6 + 6\), which is not how \(3^2 + 3^1\) should be computed.

While a more lenient educator might credit the student for arriving at the correct answer, the model chose to rate the response as wrong due to incorrect solution steps. Similarly, the model chose to mark wrong responses that contained Unicode analogs of the expected answer (e.g., À instead of A), badly formatted numbers (e.g., 9.34.5 instead of 9.345), and similar small variations that human labellers tended to classify as "correct," taking into account the student's limitations when using a mobile keyboard to type an answer. Looking further into CoT, the responses reveal that the model will sometimes make a mistake in its mathematical reasoning. This can lead to the model ignoring the expected answer, resulting in an inaccurate evaluation. Table 8 provides an example of this behaviour.

Table 8. Example of LLM with CoT Prompt Containing Faulty Mathematical Reasoning

question_id:	G6.N1.2.2.1.7
question_text:	What is the value of 7 in 4,170,348? A. 7,000,000 B. 7,000 C. 700,000 D. 70,000
expected_answer:	D
student_response:	d
human_grade:	correct_answer
llm_cot_grade:	wrong_answer
llm_cot_rationale:	We need to determine the place value of the digit 7 in the number 4,170,348. 1. The digit 7 is in the hundred thousand places. 2. The value of a digit in the hundred thousand place is the digit multiplied by 100,000. 3. Therefore, the value of 7 in 4,170,348 is 7 * 100,000 = 700,000. The correct answer is C (700,000). The student's answer is d, which corresponds to 70,000.

5.3.1. Performance vs. Latency

Table 9 shows the average and longest processing times each evaluation took to make a prediction. While CoT prompting resulted in small but stable improvements over the string processing and symbolic evaluations, it also significantly increased

response latency. On average, CoT responses took 2.79 seconds, compared to 0.73 seconds for few-shot LLM calls. The few-shot evaluation took slightly longer than the zero-shot approach. Text processing evaluations took considerably less time than all prompt-based approaches, which is expected given that this approach did not require connection to the model over the internet or the execution of a large-scale machine learning model.

Table 9. Latency of Four Answer Evaluation Approaches on 2-Class Task in Seconds

	Average Processing Time	Longest Processing Time
Text Processing	*0.006*	*0.269*
LLM Zero-shot	*0.68*	*5.687*
LLM Few-shot	*0.73*	*5.937*
LLM Chain-of-thought	*2.79*	*16.281*

These results indicate that LLM processing time can be affected by the amount of input tokens the model needs to consume in the case of a longer prompt (such as in a few-shot prompts), and can be increased significantly when the model needs to generate a significant amount of output tokens (such as in the case of chain-of-thought prompting). Additionally, prompt-based approaches could experience more fluctuation in processing time. String processing and symbolic evaluation, while less flexible and less accurate ones have much lower latency and more consistent processing time.

5.3.2. Model Reliability

While deterministic approaches like text processing provide consistent results, generative LLMs produce their output using probabilistic methods, and therefore can return different outputs given the same inputs. This variation may occur even when the temperature is set to 0. In some respects, this is similar to human raters, who occasionally will award different ratings to the same student response, when asked to re-rate it after a period of time. Measures of intra-rater reliability are intended to evaluate the extent to which a single rater agrees with their own judgment over time. To investigate the consistency of prompt-based methods, zero-shot and CoT approaches were rerun 10 times on a smaller dataset of 100 examples. As shown earlier, these two approaches scored the highest of the prompt-based approaches. For each run, the model labels were compared against the predicted labels to get a Cohen's Kappa score to measure inter-rater reliability for the run. All runs were then compared against one other to arrive at a Fleiss Kappa to represent inter-run reliability. Table 10 shows the results of these runs.

Table 10. Intra-Rater Reliability: Per-Run Agreement with Human Labels and Inter-Run Consistency

	Per-Run Agreement with Human Labels (Cohen's Kappa)										Intra-Run
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Fleiss's Kappa
LLM Zero-shot	*0.66*	*0.66*	0.68	0.70	0.66	0.62	0.66	0.70	0.66	0.66	0.90
LLM Chain-of-thought	*0.86*	*0.72*	0.74	0.74	0.74	0.72	0.74	0.66	0.70	0.72	0.88

Both CoT and zero-shot approaches had relatively high inter-run reliability as measured by Fleiss Kappa. However, the results indicate that CoT grading, while showing higher answer validity (represented by higher agreement with human labellers), has lower reliability between individual runs, and in one case scores worse than zero-shot prompting. This suggests that CoT prompting may experience more variation in how it scores responses, which may stem from its reasoning differing between runs. This could lead to accepting answers with typographical errors or other discrepancies outlined earlier, while rejecting them in other instances. While a student may not answer the same question multiple times, this variation could cause student confusion when the LLM does not consistently handle a particular answer pattern (such as substituting Unicode characters).

6. Experiment 2: Impact of Improved Grading on Student Ability Estimates

While improving model performance in grading short answer questions is a key area of research, we also seek to better understand the impact of such models on the analysis of student learning. In our second experiment, we investigated whether improved accuracy in model grading corresponded to changes in our estimates of student ability.

6.1. Experimental Design

To quantify the effect of different automated grading algorithms on predicting individual student mastery, we apply the algorithms described in the previous section to generate answer correctness labels for the entire dataset. We exclude questions labelled by human annotators as "other," as there are no straightforward ways to incorporate student non-attempts into the BKT evaluation.

Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1994) is one of the most widely used algorithms to model student knowledge in ITS (Abdelrahman et al., 2023). For any given skill, BKT assumes that a student either does or does not know it. Every time a student attempts the skill, the probability of them knowing the skill is updated based on their performance up to that point and whether they were able to demonstrate the skill correctly. Standard BKT uses four parameters to model student knowledge. Two parameters are related to learner knowledge. When first attempting to demonstrate a skill, a student has the initial probability $P(L_0)$ of knowing the skill. This probability is updated each time the student attempts to demonstrate the skill (i.e., after t attempts, the probability of knowing the skill is $P(L_t)$). At each practice opportunity, a student has a probability P(T) of learning the skill. The other two BKT parameters are related to learner performance. The probability of a student knowing the skill and yet making a mistake when attempting to demonstrate the skill is P(S). P(G) represents the probability of a student correctly guessing the answer even when not knowing the skill.

We calculated BKT scores for each student on every lesson they attempted, using only their first attempts to respond to each question. To calculate these scores, we used the following default parameters for every lesson, as suggested by Nguyen et al. (2020): P(L0)=0.4, P(T)=0.05, P(S)=0.299, and P(G)=0.299. To determine if a student had mastered a lesson, we used the last BKT score calculated for that student in each lesson. While mastery thresholds for BKT scores vary between sources, we chose a threshold of 0.9 to signify that a student had mastered the lesson. This specific threshold for mastery was determined based on previous experiments, rather than theoretically, but could also plausibly be set lower or higher.

Next, to investigate the effect of grading mechanisms on evaluating individual student mastery, we calculated the number of lessons each student mastered according to different grading algorithms. We then compared these numbers between the worst-performing algorithm (naive string matching) and the best-performing algorithm (CoT), using human labels of the student responses as the gold standard.

6.2. Results

When comparing the number of lessons that reach our threshold for mastery (BKT score of 0.9) according to different grading approaches, we find that 6.9% (165 out of 2,388) of students had their knowledge states incorrectly estimated by the baseline text processing approach. In contrast, the most successful grading approach, LLM CoT grading, only underestimated lesson mastery for 2.6% (61 out of 2,388) of students. This difference is illustrated in Table 11, which shows the effect of the grading approach by looking at a specific lesson, G7.N3.2.2.2. This lesson deals with changing forms and asks the student to present a given decimal number as a fraction. As there are multiple correct answers to this question and string-matching evaluation struggles with identifying equivalent fractions, the string-matching algorithm would regularly grade mathematically correct results as wrong.

Table 11. Change in BKT Score on Lesson G7.N3.2.2.2 by Grading Method for Example Students

user_id	BKT Estimate with String Match Grading	BKT Estimate with LLM CoT Grading	BKT Estimate with Human Grading
996	0.349435	0.845858	0.845858
1165	0.629638	0.966567	0.966567
1235	0.173999	0.809262	0.809262
1239	0.895698	0.973051	0.973051
1841	0.128321	0.913219	0.913219
2037	0.295264	0.994347	0.994347

Anecdotally, we observe that while the overall number of mis-graded responses by simpler methods like string-matching was relatively small, these errors tended to be concentrated around certain students or specific lessons. Additionally, certain lessons that allowed for multiple correct answer formats or required understanding of equivalent expressions — such as fractions — seemed to be more susceptible to grading errors from simpler methods. For one student, #1190, using string-matching to grade their answers resulted in BKT estimating that they mastered zero lessons, while both human and LLM-based grading resulted in a BKT estimate of over 0.90 for all the lessons they completed.

Another interesting case demonstrates the impact of inaccurate grading on both student experience and behaviour, as well as mastery estimation. Student #994 began their practice with multiple-choice questions in lesson G6.N1.3.6.1. However,

because they were not following the expected answer format, their correct answers were graded as wrong. This presumably caused the student to abandon the lesson midway and start a different lesson, where the situation repeated itself. The student then switched to another lesson again after just three questions. However, once they started a lesson where the answer format was less ambiguous, the grading quality improved. From that point on, not only did the student start completing the lessons, solving all 10 questions, but the estimation of their mastery also became more aligned with their actual performance.

7. Discussion and Conclusion

In this paper, we make two contributions to the field of automated short answer grading (ASAG) and learning analytics. First, through AMMORE, we expand the landscape of publicly available datasets, particularly in representing underserved educational contexts. The collection of natural language interactions from a chat-based learning environment enables research into how students engage with and learn from these increasingly common interfaces, and by providing data from West African students, AMMORE enables investigation of learning patterns and knowledge construction across different cultural contexts. Second, our experimental results demonstrate both methodological advances in interpreting student interactions with the learning platform and improvements in tracking student learning. More specifically, we find that augmenting traditional NLP methods with an LLM-driven approach allows us to improve our ability to correctly interpret the long tail of difficult-to-grade responses. This improved interpretation, in turn, can lead to significant changes in the estimation of student concept mastery.

7.1. Advancing ASAG Methodology

Our findings regarding LLM performance in grading fill-in math responses contribute to an ongoing evolution in ASAG approaches. Traditional rule-based and statistical methods, as documented by Burrows et al. (2015), have historically struggled with the variability of student responses. Our results align with recent work by Botelho et al. (2023), who found that deep learning approaches could effectively handle diverse mathematical expressions. However, we achieved high levels of accuracy using relatively simple prompting approaches such as CoT, and without requiring extensive training data needed in previous neural approaches (e.g., Shen et al.'s [2023] MathBERT).

This performance advantage particularly manifests in handling what Sung et al. (2019) termed "boundary cases" — responses that are semantically correct but syntactically variant. Where earlier work by Lan et al. (2015) required sophisticated mathematical language processing rules to handle these boundary cases, our LLM-based approach automatically adapts to different expression formats. This aligns with Gilardi et al.'s (2023) findings about LLMs' superior flexibility in text annotation tasks, though we extend their work to the specific domain of mathematical responses.

However, our results also reveal limitations similar to those noted by Kortemeyer (2023) in physics grading — particularly regarding occasional mathematical reasoning errors and sensitivity to prompt construction. These findings suggest that while LLMs offer significant advantages, they may best serve as part of a hybrid approach rather than a complete replacement for traditional methods, which echoes Schneider et al.'s (2024) conclusions about the current state of LLM-based grading.

7.2. Implications for Learning Analytics

The results of our experiments have important implications for the field of LA, particularly as educational technologies increasingly adopt chat-based interfaces. These interfaces generate rich streams of natural language data that capture student thinking and learning processes in unprecedented detail. However, this shift also presents substantial challenges for LA systems in processing and deriving meaningful insights from these interactions.

Traditional text-processing and NLP approaches, while computationally efficient and deterministic, struggle with the long tail of unexpected but valid student responses. Previous research has shown that building rule-based systems to handle this variability requires extensive engineering effort and domain expertise (Burrows et al., 2015; Lan et al., 2015), making it impractical for many learning platforms. This limitation has influenced system design significantly. Where earlier platforms often defaulted to multiple-choice questions for reliability, our results suggest that open-response questions can be reliably graded at scale, supporting Magliano and Graesser's (2012) arguments for their pedagogical value.

The superior performance of LLM-based approaches with CoT prompting suggests a promising direction for handling this variability in student interactions. The success of our approach in handling diverse response formats is especially relevant given Johnson and Green's (2006) findings about the importance of allowing multiple answer representations in mathematics assessment. Moreover, our work extends recent findings by Gurung et al. (2024) comparing multiple-choice and fill-in problems. While they focused on learning outcomes, our results address the practical implementation challenges that have historically limited the use of open-response questions in digital platforms.

Our results also complement work by Cukurova et al. (2022) on the quality of online tutoring interactions. Where they focused on process metrics, our findings demonstrate how improved response interpretation can enhance outcome measurements. The substantial reduction in misclassified mastery states (from 6.9% to 2.6%) supports Motz et al.'s (2023) argument for focusing LA research more directly on learning outcomes in applied systems. Our findings suggest that improving

the accuracy of response interpretation, even for a relatively small subset of challenging cases, can lead to significantly better estimates of student knowledge states. This improvement, in turn, enables more accurate modelling of learning trajectories.

7.3. Limitations and Further Research

Despite the promising results, our study has several limitations that suggest directions for future research. First, our dataset is limited to middle school mathematics questions from specific domains ("Algebra" and "Numbers and Operations"). Given Crossley et al.'s (2019) findings about domain-specific variations in automated assessment performance, future work should investigate how these approaches generalize across subject areas, complexity levels, and age groups. Second, our experiments focused on a binary classification of answers as correct or incorrect. This simplification, while useful for our analysis, does not capture the full spectrum of partial understanding that students may demonstrate. Future research could explore how LLM-based approaches might support more nuanced scoring rubrics, building on work by Mayfield and Black (2020) on automated essay scoring. Third, while LLMs offer impressive flexibility in handling unexpected inputs, their computational and financial costs at scale, combined with occasional unpredictability through hallucination and mathematical reasoning errors, suggest the need for research into robust verification methods, perhaps building on recent work in LLM output validation (Henkel et al., 2024). Finally, our findings indicate that learning systems will likely need to adopt hybrid approaches: using well-understood, deterministic NLP methods for common interaction patterns while reserving LLMs for handling edge cases. More research is needed on optimal architectures for such systems, including methods for automatically determining when to use each approach based on input characteristics, extending work by Allen et al. (2014) on adaptive assessment strategies.

7.4. Conclusion

Our work demonstrates both the potential and current limitations of LLM-based approaches in educational technology. While LLMs offer powerful capabilities for handling unexpected student interactions, their effective integration requires careful consideration of practical constraints and strategic deployment alongside traditional methods. As educational platforms increasingly generate rich natural language data, the LA community must continue developing scalable approaches for extracting meaningful insights about learning processes.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

No direct funding was received for this research. Author 1 has an ongoing research partnership with Rising Academies and works as a consultant on a project related to developing a conversational agent to support students' math skills. Author 2 works for Rising Academies as Research and Assessment Manager. Authors 3 and 4 work as part time consultants for Rising Academies.

Acknowledgements

We would like to thank John Whitmer and Alexis Andres for their support in assembling the AMMORE dataset. We would also like to thank Ryan Baker for guidance on BKT modelling.

References

Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. *ACM Computing Surveys, 55*(11), 224. https://doi.org/10.1145/3569576

Allen, L. K., Snow, E. L., Crossley, S. A., Tanner Jackson, G., & McNamara, D. S. (2014). Reading comprehension components and their relation to writing. *L'Année Psychologique, 114*(4), 663–691. https://doi.org/10.4074/S0003503314004047

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27*(1), 3–23. https://doi.org/10.2307/3315487

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Botelho, A., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning, 39*(3), 823–840. https://doi.org/10.1111/jcal.12793

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Cechinel, C., Ochoa, X., Lemos Dos Santos, H., Carvalho Nunes, J. B., Rodés, V., & Marques Queiroga, E. (2020). Mapping learning analytics initiatives in Latin America. *British Journal of Educational Technology, 51*(4), 892–914. https://doi.org/10.1111/bjet.12941

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications, 40*(11), 4715–4729. https://doi.org/10.1016/j.eswa.2013.02.007

Cochran, K., Cohn, C., Hutchins, N., Biswas, G., & Hastings, P. (2022). Improving automated evaluation of formative assessments with text data augmentation. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education: 23rd international conference, AIED 2022, Durham, UK, July 27–31, 2022, proceedings, part I* (pp. 390–401). Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_32

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278. https://doi.org/10.1007/BF01099821

Crossley, S. A., Kim, M., Allen, L., & McNamara, D. (2019). Automated summarization evaluation (ASE) using natural language processing tools. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial intelligence in education: 20th international conference, AIED 2019, Chicago, IL, USA, June 25–29, 2019, proceedings, part I* (pp. 84–95). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_8

Cukurova, M., Khan-Galaria, M., Millán, E., & Luckin, R. (2022). A learning analytics approach to monitoring the quality of online one-to-one tutoring. *Journal of Learning Analytics, 9*(2), 105–120. https://doi.org/10.18608/jla.2022.7411

Dey, I., Gnesdilow, D., Passonneau, R., & Puntambekar, S. (2024). Potential pitfalls of false positives. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education: Posters and late-breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky: 25th international conference, AIED 2024, Recife, Brazil, July 8–12, 2024, proceedings, part I* (pp. 469–476). Springer Cham. https://doi.org/10.1007/978-3-031-64315-6_45

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction, 19*(3), 243–266. https://doi.org/10.1007/s11257-009-9063-7

Funk, S. C., & Dickson, K. L. (2011). Multiple-choice and short-answer exam performance in a college classroom. *Teaching of Psychology, 38*(4), 273–277. https://doi.org/10.1177/0098628311421329

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education, 57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Gurung, A., Vanacore, K., McReynolds, A. A., Ostrow, K. S., Worden, E., Sales, A. C., & Heffernan, N. T. (2024). Multiple choice vs. fill-in problems: The trade-off between scalability and learning. *Proceedings of the 14th Learning Analytics and Knowledge Conference* (LAK '24), 18–22 March 2024, Kyoto, Japan (pp. 507–517). https://doi.org/10.1145/3636555.3636908

Hahn, M. G., Navarro, S. M. B., De La Fuente Valentín, L., & Burgos, D. (2021). A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access, 9*, 108190–108198. https://doi.org/10.1109/ACCESS.2021.3100890

Henkel, O. (2024, March 21). *Rori - Quick intro* [Video]. YouTube. https://www.youtube.com/watch?v=xXg6XRajbbk

Henkel, O., Hills, L., Roberts, B., & McGrane, J. (2024). Can LLMs grade open response reading comprehension questions? An empirical study using the ROARs dataset. *International Journal of Artificial Intelligence in Education.* https://doi.org/10.1007/s40593-024-00431-z

Hsu, S., Li, T. W., Zhang, Z., Fowler, M., Zilles, C., & Karahalios, K. (2021). Attitudes surrounding an imperfect AI autograder. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 8–13 May 2021, Yokohama, Japan (Article 681). https://doi.org/10.1145/3411764.3445424

Injeti, A. S., Rupsica, G. N., Reddy, G. P., Balakrishnan, R. M., & Pati, P. B. (2024). A machine learning-based classification of students' algebraic responses using MathBERT embeddings. *Proceedings of the 2024 5th International Conference for Emerging Technology* (INCET), 24–26 May 2024, Belgaum, India (pp. 1–6). https://doi.org/10.1109/INCET61516.2024.10593432

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *The Journal of Technology, Learning and Assessment, 4*(5). https://ejournals.bc.edu/index.php/jtla/article/view/1652

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large language models are zero-shot reasoners.* arXiv. https://doi.org/10.48550/arXiv.2205.11916

Kortemeyer, G. (2023). *Performance of the pre-trained large language model GPT-4 on automated short answer grading*. arXiv. https://doi.org/10.48550/arXiv.2309.09338

Lan, A. S., Vats, D., Waters, A. E., & Baraniuk, R. G. (2015). Mathematical language processing: Automatic grading and feedback for open response mathematical questions. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (L@S '15), 14–18 March 2015, Vancouver, BC, Canada (pp. 167–176). https://doi.org/10.1145/2724660.2724664

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods, 44*(3), 608–621. https://doi.org/10.3758/s13428-012-0211-3

Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 10 July 2020, Seattle, WA, USA (pp. 151–162). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.15

Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical program. *International Medical Education, 3*(1), 32–43. https://doi.org/10.3390/ime3010004

Motz, B. A., Bergner, Y., Brooks, C. A., Gladden, A., Gray, G., Lang, C., Li, W., Marmolejo-Ramos, F., & Quick, J. D. (2023). A LAK of direction: Misalignment between the goals of learning analytics and its research scholarship. *Journal of Learning Analytics*, *10*(2), 1–13. https://doi.org/10.18608/jla.2023.7913

Nguyen, H. A., Hou, X., Stamper, J., McLaren, B. M. (2020). Moving beyond test scores: Analyzing the effectiveness of a digital learning game through learning analytics. *Proceedings of the 13th International Conference on Educational Data Mining* (EDM 2020), 10–13 July 2020, Online (pp. 487–495). International Educational Data Mining Society.

O'Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 11*(4), 331–351. https://doi.org/10.1207/s15324818ame1104_3

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. https://doi.org/10.48550/arXiv.2203.02155

Pelánek, R. (2015). Metrics for evaluation of student models. *Journal of Educational Data Mining, 7*(2), 1–19. https://doi.org/10.5281/zenodo.3554665

Pulman, S. G., & Sukkarieh, J. Z. (2005). Automatic short answer marking. *Proceedings of the Second Workshop on Building Educational Applications Using NLP* (EdAppsNLP 05), 29 June 2005, Ann Arbor, MI, USA (pp. 9–16). Association for Computational Linguistics https://doi.org/10.3115/1609829.1609831

Rajendran, R., Iyer, S., & Murthy, S. (2019). Personalized affective feedback to address students' frustration in ITS. *IEEE Transactions on Learning Technologies, 12*(1), 87–97. https://doi.org/10.1109/TLT.2018.2807447

Rising Academies. (2024). Rori [Software]. Available from https://rori.ai/

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441–474. https://doi.org/10.1191/0265532206lt337oa

Schneider, J., Schenk, B., & Niklaus, C. (2024). Towards LLM-based autograding for short textual answers. *Proceedings of the 16th International Conference on Computer Supported Education* (CSEDU 2024), 2–4 May 2024, Angers, France (pp. 280–288). SciTePress. https://doi.org/10.5220/0012552200003693

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., Graff, B., & Lee, D. (2023). *MathBERT: A pre-trained language model for general NLP tasks in mathematics education.* arXiv. https://doi.org/10.48550/arXiv.2106.07340

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2022). *Learning to summarize from human feedback.* arXiv. https://doi.org/10.48550/arXiv.2009.01325

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), 3–7 November 2019, Hong Kong, China (pp. 6070–6074). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1628

USAID. (2019). *Global proficiency framework: Reading and mathematics.* United States Agency for International Development. https://www.edu-links.org/resources/global-proficiency-frameworkreading-and-mathematics

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models*. arXiv. https://doi.org/10.48550/arXiv.2206.07682