Leveraging Large Language Models for Evaluating Explanations in Math Education

ANONYMOUS AUTHOR(S)

Human-conducted rating tasks are resource-intensive and demand significant time and financial commitments. As Large Language Models (LLMs) like GPT emerge and exhibit prowess across various domains, their potential in automating such evaluation tasks becomes evident. In this research, we leveraged four prominent LLMs: GPT-4, GPT-3.5, Vicuna, and PaLM 2, to scrutinize their aptitude in evaluating teacher-authored mathematical explanations. We utilized a detailed rubric that encompassed accuracy, explanation clarity, the correctness of mathematical notation, and the efficacy of problem-solving strategies. During our investigation, we unexpectedly discerned the influence of HTML formatting on these evaluations. Notably, GPT-4 consistently favored explanations formatted with HTML, whereas the other models displayed mixed inclinations. When gauging Inter-Rater Reliability (IRR) among these models, only Vicuna and PaLM 2 demonstrated high IRR using the conventional Cohen's Kappa metric for explanations formatted with HTML. Intriguingly, when a more relaxed version of the metric was applied, all model pairings showcased robust agreement. These revelations not only underscore the potential of LLMs in providing feedback on student-generated content but also illuminate new avenues, such as reinforcement learning, which can harness the consistent feedback from these models.

CCS Concepts: • **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**; **E-learning**.

Additional Key Words and Phrases: Mathematics Education, Teacher-authored Explanations, Large Language Models, LLMs, GPT, PaLM, Vicuna, Evaluation, Inter-rater Reliability, Educational Technology

ACM Reference Format:

Anonymous Author(s). 2024. Leveraging Large Language Models for Evaluating Explanations in Math Education. In . ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Mathematics education serves as a pivotal element in fostering critical thinking and problem-solving skills among students. High-quality explanations in this field are not merely an academic requirement; they are a crucial factor in shaping students' educational experience, engagement, and performance. However, creating and evaluating explanations for math problems can be a time-consuming and costly endeavor.

Teachers have traditionally been the primary medium for the transfer of mathematical knowledge, adapting their methods to fit diverse learning styles and needs. In recent years, advancements in educational technology have offered supplementary avenues for student support. Among these technologies, Large Language Models (LLMs), such as GPT-3.5-turbo and GPT-4, have emerged as promising tools for automating the evaluation and potential improvement of educational content. As educational platforms increasingly integrate these advanced technologies, there is a timely need to assess their effectiveness rigorously.

While the existing body of work provides valuable insights, there are noticeable gaps and challenges that this study aims to address. Firstly, much of the research has traditionally focused on textbook or professionally crafted explanations

Manuscript submitted to ACM

rather than teacher-generated content [16]. There is a lack of understanding of the nature and quality of teacher-authored math explanations for supporting students math learning. This often overlooks the unique challenges that teachers face in authoring high-quality explanations, such as time constraints and the need for specialized knowledge in certain topics [10]. Addressing these challenges is pivotal, as it lays the groundwork for the necessity of LLMs in this context. By automating the evaluation and potential improvement of explanations, LLMs can alleviate the time constraint burden on teachers and provide insights that contribute to specialized knowledge, thereby enhancing the quality of teacher-authored content. Secondly, while there are diverse rubrics that assess students' explanations on open response problems that require students to explain their reasoning, manually assessing a large volume of explanations is cumbersome and time-consuming. The emergence of LLMs provides a great potential to evaluate teacher-authored explanations on a large scale, however, there is no universally agreed-upon set of criteria for evaluating the quality explanations [20], which makes it difficult to efficiently leverage LLMs to evaluate explanations.

Our study aims to fill these gaps by focusing on real-world, teacher-authored explanations and employing a robust, multi-faceted rubric for their evaluation. We also leverage state-of-the-art LLMs to explore how AI can assist in evaluating math explanations.

1.1 Purpose and Research Questions

This study, which we have pre-registered through the Open Science Foundation[1], aims to use LLMs to evaluate teacher-authored explanations for math problems. We will compare the capabilities of LLMs including GPT-3.5-turbo, GPT-4, PaLM 2 [2] and Vicuna [4]. The primary research questions for this study are:

RQ1 How can LLMs be used to *evaluate* teacher-authored explanations in math education?

RQ2 What is the inter-rater reliability when using multiple LLMs to evaluate the explanations?

Our study provides several contributions for educators and researchers alike. Central to our work is a curated dataset of 853 teacher-authored explanations, encompassing both HTML-formatted and non-formatted versions. To ensure a systematic evaluation of these explanations, we put forth a rubric that not only sets a benchmark for our evaluation but also serves the broader educational community in assessing the caliber of mathematical explanations. Furthering our study's depth, we engaged in evaluations employing state-of-the-art LLMs, which we also make available. Our analysis looks at each LLM and hints at their potential in online math education scenarios for evaluating teacher-authored explanations. This approach underscores our commitment to integrating LLMs into the dynamic landscape of online mathematics education.

2 LITERATURE REVIEW

A rich body of literature exists that discusses the importance of explanations in educational settings, particularly in the realm of math education. The following subsections delve into various dimensions of this topic.

2.1 Explanations in Math Education

Explanations are pedagogical actions that provide solutions or answers to questions posed by teachers or students [22]. In math education, explanations could provide students with arithmetic or algorithmic procedures for solving math problems. Explanations are a common and direct way of instruction. Explanations can provide students with insightful and straightforward ways to understand mathematical concepts and principles. Although students with higher prior

[1]The registration has been embargoed during the double-blind review process, but will be made available upon acceptance.

knowledge may benefit from instructions that are more cognitively demanding [13], complete and direct step-by-step explanations might cause less cognitive load and could be more beneficial for novice or lower-performing students to build foundational understanding of math concepts or principles [7].

While explanations are potentially helpful for students, they have to be carefully crafted. The quality of explanations is a major factor that impacts learning effectiveness [22]. Inaccurate and unclear explanations could be harmful to students. It is important to provide accurate and clear explanations so students will not be misled or get confused. Furthermore, explanations should focus on central concepts to support students' conceptual understanding and help students transfer the problem solving procedures to other math problems [18, 22]. Finally, explanations need to provide complete step-by-step procedures as the completeness of explanations is critical for students' learning gains [5, 8, 17]. In summary, math explanations should be accurate and clear, and use complete and appropriate problem-solving strategies.

2.2 Teacher-Authored Math Explanations

Research into teacher-authored explanations for mathematical problems is a sub-field that builds upon existing studies on effective teaching methods and learning outcomes [6]. The effectiveness of teacher-generated solutions is linked to fostering conceptual understanding [11]. Moreover, the significance of delivering clear and concise explanations is a cornerstone of effective teaching, corroborated by the principles of instructional design [12].

Teacher-authored math explanations serve as a primary source of knowledge for students and have a profound influence on their educational experience [13]. The quality and clarity of math explanations are often associated with improved student learning outcomes. Research found that process-oriented math explanations with clarification on why a certain step in the solution was required has a more significant positive impact on learning gains than explanations with solutions steps without the clarification and students in both conditions had significantly higher learning gains than students who received just the problem without explanations [8]. Providing high-quality explanations is essential to support students' math learning [20].

In enhancing the quality and precision of these teacher-authored math explanations, the utilization of instructional rubrics can be instrumental. As shown by Andrade [1], these rubrics do more than just evaluate: they can also be teaching tools that underline the expected quality in explanations. For instance, a rubric can detail criteria such as the clarity of a mathematical concept, the logic behind each step, and the relevancy of examples provided. Through self-assessment using these rubrics, teachers can critically review and subsequently refine their explanations to align better with instructional best practices and student needs. By merging the realms of instruction and assessment with rubrics, teacher-authored math explanations can achieve a balance between content depth and instructional clarity.

2.3 Leveraging LLMs to Evaluate Math Explanations

Machine learning algorithms have been instrumental in generating and assessing educational content, with LLMs such as the GPT-x series being at the forefront of this innovation. Most research has been focusing on using LLMs to generate content. A study by Prihar et al. [15] explores the capabilities of GPT-3 in authoring explanations for middle-school mathematics problems. The goal of this exploration is to leverage the model's potential to quickly integrate explanations for emerging mathematics curricula into online learning platforms.

In their methodology, Prihar et al. employed two approaches. The first summarized the essential advice from tutoring chat logs between students and live tutors. The second attempted to generate explanations through few-shot learning from explanations provided by teachers for similar mathematics problems. A comparative survey revealed that while the synthetic explanations generated by GPT-3 could not outperform those authored by teachers, there is potential for

more powerful future models to be employed effectively. The authors suggest that GPT-3 and subsequent models can serve as valuable tools to augment teachers' process of writing explanations, rather than replacing them.

While LLMs have shown potential in generating educational content, using LLMs to evaluate math explanations remains a new and promising area for further exploration. It is pivotal to continue with research and development in this area to harness the full potential of LLMs in education.

2.4 Theoretical Foundations Underpinning the Rubric

The rubric to assess explanations to open-ended responses in mathematics is aligned with several theoretical underpinnings, deeply rooted in educational psychology and pedagogical research. This section provides an overview of the primary theoretical frameworks that underlie the structure and elements of the rubric shown in Section 3.3.1.

2.4.1 Constructivist Foundations. At the heart of assessing mathematical understanding lies the constructivist theory, emphasizing the active role of learners in constructing knowledge. Piaget's work offers profound insights into how individuals build and refine cognitive structures in response to experiences [14]. In evaluating the "accuracy" and "problem-solving strategy" in the rubric, the focus is on gauging how well the explanations have constructed and internalized mathematical concepts and processes.

2.4.2 Bloom's Taxonomy and Cognitive Development. Bloom's Taxonomy offers a hierarchical framework of cognitive objectives, ranging from knowledge recall to higher-order thinking skills such as synthesis and evaluation [3]. Elements of the rubric, such as "accuracy" and "problem-solving strategy," directly map to various levels of this taxonomy, capturing both the depth and breadth of understanding.

2.4.3 Socio-cultural Theory and Mathematical Communication. The importance of using appropriate "mathematical language and notation" in the rubric is deeply rooted in Vygotsky's socio-cultural theory [21]. This theory accentuates that learning is a culturally mediated process. Rogoff's exploration of cognitive development in a social context further strengthens this perspective [19]. By emphasizing the correct use of mathematical language, the rubric inherently respects the cultural norms of the mathematical community, ensuring the explanations align with established conventions.

In summary, the rubric, while a practical tool for assessment, is deeply entrenched in several foundational educational theories, ensuring its effectiveness and relevance in diverse mathematical learning contexts.

3 METHODOLOGY

3.1 Data Collection

Our dataset consists of 853 teacher-authored text-based explanations corresponding to various math problems across multiple grade levels. The majority of the problems are from Grades 6-8, with one problem each from Grade 4 and Algebra 2. The specific distribution of explanations across different grade levels is detailed in Table 1. The explanations were created for open response problems that ask students to explain their reasoning or solution of math problems.

3.2 Data Processing and Challenges

In our initial data collection phase, we aimed to employ LLMs to evaluate teacher-authored explanations for mathematical problems. These explanations, when rendered on our online learning platform, contained HTML formatting tags, which provided essential context for the problems and explanations. This included, but was not limited to, the use of MathJax

Grade Level	# Explanations
Grade 4	1
Grade 6	356
Grade 6 Accelerated	72
Grade 7	202
Grade 7 Accelerated	84
Grade 8	137
Algebra 2	1

Table 1. Distribution of Explanations by Grade Level

to accurately display mathematical notations. In an oversight, one member of our research team removed the HTML tags during the data preprocessing phase, intending to simplify the explanations for the LLMs, while another researcher was under the assumption that these tags were still present and anticipated the LLM-generated explanations would mirror this format.

This discrepancy in understanding and processing had potential ramifications for the LLMs' interpretation and evaluation of the explanations. Without the HTML and MathJax formatting, there was a substantial risk of decreased fidelity in the LLM evaluations, making them potentially unrepresentative of the original teacher-authored explanations' intended format and meaning.

Given these concerns and the intrinsic value of the original formatting for accurate evaluation, we decided to narrow the scope of our study. We undertook a second round of data collection, ensuring the inclusion of the original HTML formatting in the problems and explanations. Our revised focus centered on evaluating teacher-authored explanations in their intended format.

Due to time constraints and the depth of analysis required for a thorough evaluation, the creation of enhanced explanations based on the LLM evaluations and rubric scores will be earmarked for future work. However, in the ensuing sections, we will attempt to shed light on any noticeable differences in LLM evaluations between the datasets with and without HTML tags.

3.3 Using LLMs for Explanation Evaluation

We employed four LLMs: GPT-3.5-turbo, GPT-4, PaLM 2, and Vicuna. Initially, these models were envisioned to serve two primary roles:

- **Evaluation:** The LLMs were used to rate teacher-authored explanations according to a predefined rubric.
- **Improvement:** Based on the rubric scores, it was anticipated that the LLMs would generate improved versions of the explanations.

Due to the challenges encountered during data processing, particularly the inadvertent removal and subsequent reinclusion of HTML tags, and the constraints of time, our primary focus shifted predominantly towards evaluation. While the potential for using LLMs for improvement remains promising, this aspect has been earmarked for future exploration.

3.3.1 Evaluation Rubric. To assess teacher-authored explanations, we developed a rubric with the assistance of GPT-4. The development of the rubric was guided by a specific query: "Can you suggest a rubric to evaluate a student's open-ended response in the context of math learning?". The rubric generated by GPT-4 was reviewed by experts in

learning science, math education, and computer science to ensure the rubric's validity and applicability. Furthermore, the rubric is aligned with the theoretical foundations described earlier in this paper in Section 2.4.

This rubric captures the quality of the explanations along four dimensions: Accuracy, Explanation Clarity, Mathematical Language and Notation, and Problem-Solving Strategy. Each dimension is on a scale from 0 to 4, allowing for a detailed and nuanced evaluation.

The rubric used is detailed as follows:

Accuracy (0-4 points):

0: The response is incorrect or irrelevant to the question.
1: The response contains some correct elements but has major inaccuracies.
2: The response is partially correct but contains some errors or misconceptions.
3: The response is mostly correct but has minor errors or lacks precision.
4: The response is entirely correct and demonstrates a complete understanding of the problem.

Explanation Clarity (0-4 points):

0: No explanation provided or the explanation is incomprehensible.
1: The explanation is difficult to follow and lacks coherence.
2: The explanation is somewhat clear but could be better organized or more concise.
3: The explanation is mostly clear and easy to follow but has minor issues with organization or clarity.
4: The explanation is clear, concise, and well-organized, making it easy to understand the student's thought process.

Mathematical Language and Notation (0-4 points):

0: No mathematical language or notation is used or used inappropriately.
1: Limited use of mathematical language or notation, with significant errors or inconsistencies.
2: Some correct use of mathematical language and notation, but with occasional errors or inconsistencies.
3: Mostly correct use of mathematical language and notation, with only minor errors or inconsistencies.
4: Appropriate and accurate use of mathematical language and notation throughout the response.

Problem-Solving Strategy (0-4 points):

0: No problem-solving strategy is evident or the strategy used is irrelevant or inappropriate.
1: The problem-solving strategy used is partially correct but contains significant errors or omissions.
2: The problem-solving strategy used is mostly correct but has minor errors or lacks efficiency.
3: The problem-solving strategy used is appropriate and mostly efficient, with only minor room for improvement.
4: The problem-solving strategy used is appropriate, efficient, and demonstrates a deep understanding of the problem.

3.3.2 Prompts for Evaluation with LLMs. To operationalize our evaluation process using the LLMs, we embedded the rubric, as detailed in the preceding section, within a structured prompt. This ensured that the LLMs consistently and comprehensively applied the rubric criteria when evaluating the teacher-authored explanations. The following represents the structure of the prompt:

```
The following is your rubric:

[Refer to the detailed rubric in the Evaluation Rubric section.]
```

```
Here is the problem:
{body}

Here is the explanation:
{value}

Use the rubric to provide an accurate score and justification for the explanation using the
following JSON:
{
  "score": [
    {
      "category": "Accuracy",
      "score": 0..4,
      "justification": "Explain why the score was given for accuracy."
    },
    {
      "category": "Explanation clarity",
      "score": 0..4,
      "justification": "Explain why the score was given for explanation clarity."
    },
    {
      "category": "Mathematical language and notation",
      "score": 0..4,
      "justification": "Explain why the score was given for mathematical language and notation."
    },
    {
      "category": "Problem-solving strategy",
      "score": 0..4,
      "justification": "Explain why the score was given for problem-solving strategy."
    }
 ]
}
```

The system initiated each session by setting the role for the LLM as that of a math teacher evaluating open-ended responses to math problems. This approach was intended to orient the LLM toward the specific task at hand.

```
messages=[
{'role': 'system', 'content': 'You are a math teacher using a rubric to evaluate open-ended
responses to a variety of math problems.'},
{'role': 'user', 'content': user_prompt_use_rubric}
]
```

3.4 Inter-Rater Reliability (IRR) Among LLMs

In our study, multiple LLMs were used to evaluate teacher-authored explanations based on the predefined rubric. Ensuring consistency in evaluations across these models necessitates the measurement of the Inter-Rater Reliability (IRR) among them. The IRR quantifies the level of agreement or consistency in ratings furnished by different raters, with LLMs being our raters in this context.

To compute the IRR, we employed Cohen's Kappa (κ), a statistical coefficient that gauges the degree of agreement between two raters beyond what might be expected by sheer chance. Cohen's Kappa offers the advantage of considering accidental agreement, making it a more robust measure than straightforward percentage agreement. A κ value converging to 1 denotes strong rater agreement, whereas a value near 0 suggests an agreement merely by chance.

Alongside Cohen's Kappa, we also utilized a 'Relaxed Kappa'. Conceptually, this is akin to a Weighted Kappa, which typically assigns weights to different types of disagreements based on their severity. Unlike a strict Cohen's Kappa where only exact matches count as agreements, the Relaxed Kappa permits slight deviations (as defined by a tolerance level) between raters' assessments to be considered as agreements. The rationale behind introducing a Relaxed Kappa is to accommodate scenarios where minor discrepancies in ratings across LLMs are still deemed acceptable and should not drastically affect the reliability score. In our implementation, this tolerance parameter determines the maximum allowed difference for a response to be deemed as an agreement. For instance, with a tolerance of 1, a rating of 4 by one LLM and a rating of 3 by another would still be treated as concordant.

This dual approach, employing both Cohen's Kappa and Relaxed Kappa, provides a comprehensive understanding of the IRR, catering to both strict and lenient evaluation scenarios.

4 RESULTS

4.1 Evaluation of Teacher-Authored Explanations

The evaluation process aimed to uncover significant patterns and variations between the different LLMs for their ratings in each rubric category for the teacher-authored explanations.

Descriptive Statistics: We computed the mean, median, standard deviation, minimum, and maximum for each rubric category. The summarized results for GPT-4, GPT-3.5-turbo, PaLM 2 and Vicuna can be found in Table 2, Table 3, Table 4 and Table 5 respectively.

Table 2. Descriptive Statistics for Each Rubric Category for GPT-4, With and Without HTML

	With HTML			Without HTML		
Rubric Category	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Accuracy	3.580	4.0	0.986	3.421	4.0	1.125
Explanation Clarity	3.325	4.0	0.938	3.066	3.0	1.021
Mathematical Language	2.938	3.0	0.970	2.760	3.0	0.990
Problem-Solving Strategy	3.262	4.0	1.029	2.996	3.0	1.160

Table 3. Descriptive Statistics for Each Rubric Category for GPT-3.5-turbo, With and Without HTML

	With HTML			Without HTML		
Rubric Category	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Accuracy	3.145	3.0	0.853	3.170	3.0	0.818
Explanation Clarity	3.406	4.0	0.717	3.342	3.0	0.741
Mathematical Language	2.947	3.0	1.196	2.957	3.0	1.138
Problem-Solving Strategy	2.708	3.0	1.482	2.748	3.0	1.400

Table 4. Descriptive Statistics for Each Rubric Category for PaLM 2, With and Without HTML

	With HTML			Without HTML		
Rubric Category	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Accuracy	3.518	4.0	0.792	3.532	4.0	0.752
Explanation Clarity	3.518	4.0	0.813	3.511	4.0	0.797
Mathematical Language	3.543	4.0	0.783	3.545	4.0	0.752
Problem-Solving Strategy	3.528	4.0	0.782	3.537	4.0	0.740

Table 5. Descriptive Statistics for Each Rubric Category for Vicuna, With and Without HTML

	With HTML			Without HTML		
Rubric Category	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
Accuracy	3.564	4.0	0.709	3.448	3.0	0.500
Explanation Clarity	3.576	4.0	0.757	3.215	3.0	0.438
Mathematical Language	3.621	4.0	0.677	3.972	4.0	0.172
Problem-Solving Strategy	3.586	4.0	0.694	3.890	4.0	0.321

Influence of HTML Formatting: To explore the unexpected role of HTML formatting in our evaluation, we performed paired t-tests comparing evaluations with and without HTML for both GPT-4 and GPT-3.5-turbo. The results are presented in Table 6.

Table 6. Comparison of paired t-test results between GPT-4 and GPT-3.5-turbo for evaluations with and without HTML formatting.

	GPT-4			GPT-3.5-turbo		
Category	t-statistic	Uncorrected p-value	Corrected p-value	t-statistic	Uncorrected p-value	Corrected p-value
Accuracy	4.7986	< .001	< .001	-0.7987	0.4247	1.0000
Explanation Clarity	8.8974	< .001	< .001	2.5038	0.0125	0.0499
Math. Lang. & Notation	5.6372	< .001	< .001	-0.2701	0.7871	1.0000
Prob. Solving Strategy	7.8854	< .001	< .001	-0.8714	0.3838	1.0000

Table 7. Comparison of paired t-test results between PaLM 2 and Vicuna for evaluations with and without HTML formatting.

Category	PaLM 2			Vicuna		
	t-statistic	Uncorrected p-value	Corrected p-value	t-statistic	Uncorrected p-value	Corrected p-value
Accuracy	1.4781	0.1397	0.5589	-4.2518	< .001	< .001
Explanation Clarity	-0.6186	0.5363	1.0000	-11.9566	< .001	< .001
Math. Lang. & Notation	0.2389	0.8112	1.0000	15.0892	< .001	< .001
Prob. Solving Strategy	0.9561	0.3393	1.0000	12.0380	< .001	< .001

Frequency Tables: Apart from descriptive statistics and t-tests, we also compiled frequency tables to understand the distribution and concentration of scores across various bins. For instance, looking at the frequency table for GPT-4 presented in Table 8 shows the Accuracy category with a score of 4.0 (with HTML) was obtained 688 times.

In the GPT-4 frequency distribution table (Table 8), the highest score (4.0) was the most frequent across all categories, especially prominent in the 'Accuracy' category. Interestingly a score of 3.5 was seen a few times, although this was not intended to be a score used by the rubric. It is possible that changing the prompt used to specify only using integers would have eliminated this anomaly.

GPT-3.5-turbo (Table 9) primarily exhibited scores in the 3.0 and 4.0 range. Of particular note, the 'Problem Solving' category when evaluated with HTML had a relatively high occurrence of 0.0 scores.

PaLM 2's evaluations (Table 10) showcased a prominent trend towards the upper score range, with scores of 3.0 and 4.0 being highly prevalent. Lower scores, such as 0.0, 1.0, and 2.0, were notably less frequent across categories.

For Vicuna (Table 11), the score of 4.0 was dominant in all categories. The 'Mathematical Notation' category when evaluated without HTML had a staggering occurrence of the 4.0 score.

Table 8. Frequency Distribution Using GPT-4 (With and Without HTML)

	With HTML Scores						**Without HTML Scores**					
Category	**0.0**	**1.0**	**2.0**	**3.0**	**3.5**	**4.0**	**0.0**	**1.0**	**2.0**	**3.0**	**3.5**	**4.0**
Accuracy	30	27	49	59	0	688	42	38	69	74	0	630
Explanation Clarity	15	35	86	238	1	478	26	40	142	288	1	356
Mathematical Notation	24	32	189	335	1	272	24	64	209	350	4	202
Problem Solving	21	53	84	218	1	476	36	74	137	216	0	390

Table 9. Frequency Distribution Using GPT-3.5-turbo (With and Without HTML)

	With HTML Scores					**Without HTML Scores**				
Category	**0.0**	**1.0**	**2.0**	**3.0**	**4.0**	**0.0**	**1.0**	**2.0**	**3.0**	**4.0**
Accuracy	2	38	131	345	337	5	21	132	361	334
Explanation Clarity	5	13	47	354	434	4	13	75	356	405
Mathematical Notation	66	44	105	292	346	56	42	109	322	324
Problem Solving	143	49	92	199	370	117	47	127	205	357

Table 10. Frequency Distribution Using PaLM 2 (With and Without HTML)

	With HTML Scores					Without HTML Scores				
Category	**0.0**	**1.0**	**2.0**	**3.0**	**4.0**	**0.0**	**1.0**	**2.0**	**3.0**	**4.0**
Accuracy	22	3	20	274	534	18	1	24	276	534
Explanation Clarity	22	4	31	249	547	18	7	35	254	539
Mathematical Notation	22	3	14	265	549	19	1	18	273	542
Problem Solving	22	3	13	280	535	18	1	16	288	530

Table 11. Frequency Distribution Using Vicuna (With and Without HTML)

	With HTML Scores					Without HTML Scores				
Category	**0.0**	**1.0**	**2.0**	**3.0**	**4.0**	**0.0**	**1.0**	**2.0**	**3.0**	**4.0**
Accuracy	12	2	31	256	552	0	0	1	469	383
Explanation Clarity	12	13	29	217	582	0	0	10	650	193
Mathematical Notation	12	2	17	235	587	0	0	1	22	830
Problem Solving	12	2	24	251	564	0	0	2	90	761

Correlation Heatmaps: Correlation analysis is instrumental in discerning the strength and direction of relationships between variables. In our study, this technique illuminated how various rubric categories interrelate, revealing if high performance in one category suggests similar performance in another. Such insights help uncover latent behavioral patterns.

The visual representation of these correlations was realized through heatmaps. Within these heatmaps, each cell represents the Pearson correlation coefficient between two rubric categories. Color intensity and hue signify the strength and direction: warmer shades (e.g., red tones) denote positive correlations, implying that as one variable rises, so does the other. Conversely, cooler shades (e.g., blue tones) mark negative correlations, indicating inverse relationships between the variables.

For GPT-4, the correlation heatmaps, both with and without HTML exhibited analogous correlation coefficients. While Table 2 displayed a statistically significant variation in the rubric categories' descriptive measures, their inter-correlations remained largely unaffected. In contrast, GPT-3.5-turbo's correlation patterns echoed GPT-4's, but with two exceptions: (Accuracy x Problem Solving Strategy) and (Explanation Clarity x Problem Solving Strategy) both showed a noticeable uptick for GPT-4.

Unique to the dataset was Vicuna (without HTML), depicted in Subfigure 1h. It exhibited negative correlations, specifically between (Accuracy x Explanation Clarity) and (Problem Solving Strategy x Explanation Clarity).

In analyzing PaLM 2 (with HTML) shown in Subfigure 1e, it was evident that the rubric categories were profoundly interrelated. Such pronounced correlations suggest a concurrent change in one category usually echoed across others. PaLM 2's performance without HTML mirrored this trend, albeit with slightly diminished correlation coefficients. This observation implies that, using our methodology, PaLM 2 did not assess the rubric categories in isolation.

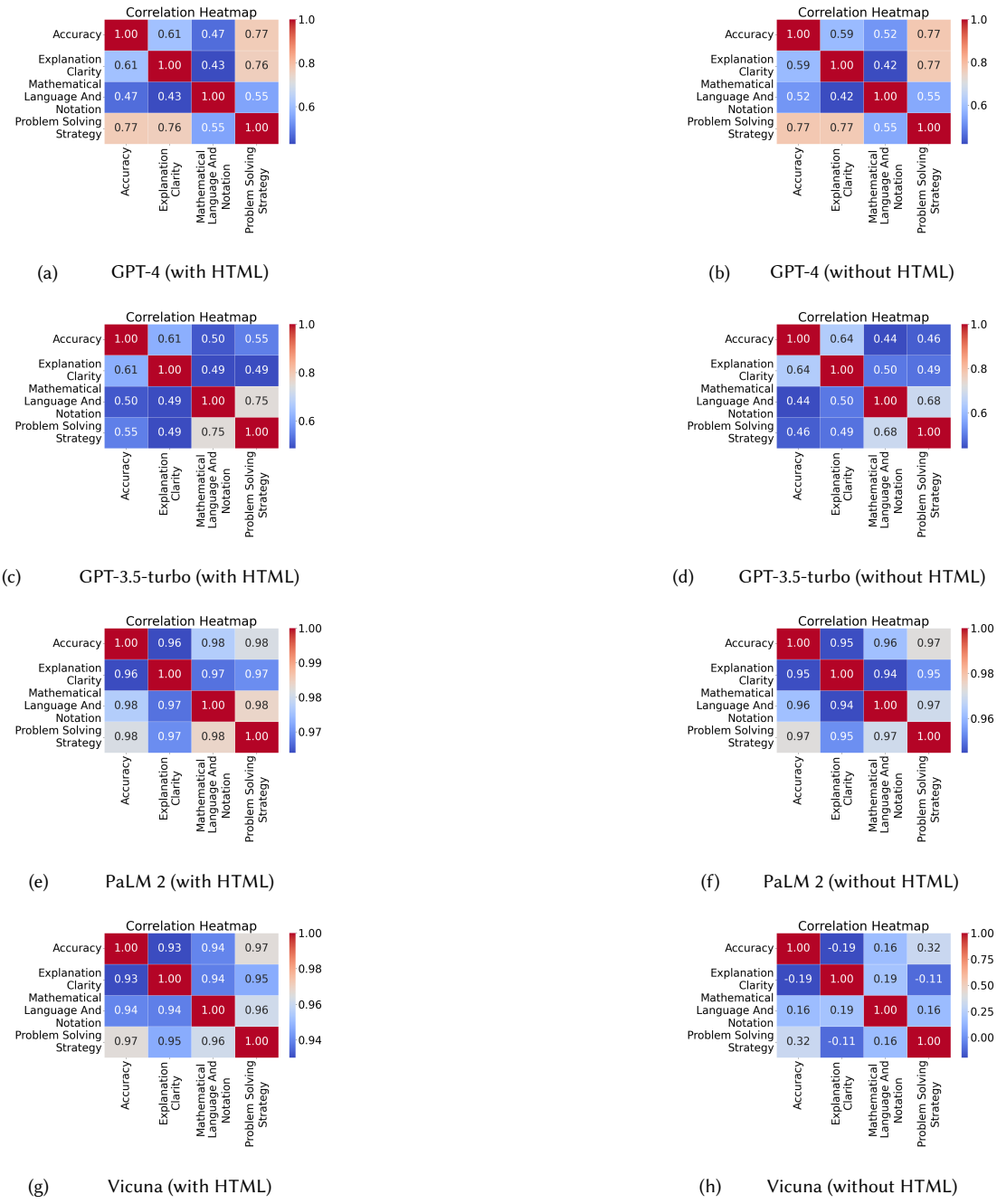

Fig. 1. Correlation heatmaps among rubric categories for each LLM used to evaluate with and without HTML tags.

4.2 Inter-rater Reliability

Inter-rater reliability measures the degree of agreement among raters. When assessing agreement with multiple raters or models, there are various established methodologies that can be utilized. Cohen's Kappa (κ) is a standard metric for two raters. However, when extending to multiple raters, Light [9] provides generalizations and alternative measures to Cohen's Kappa, especially when assessing "internal" agreement among several observers.

In our study, we followed an approach aligning with these generalizations. The inter-rater reliability between each pair of models was assessed using Cohen's Kappa for each of the four rubric categories: Accuracy, Explanation Clarity, Mathematical Language and Notation, and Problem Solving Strategy. We averaged the results between each pair of models to compute an average Cohen's Kappa. This procedure was then repeated for a relaxed Cohen's Kappa. The results are summarized in Tables 12 to 15.

Model	PaLM 2	Vicuna	GPT-3.5-turbo	GPT-4
PaLM 2	1	0.65	0.18	0.17
Vicuna	0.65	1	0.16	0.17
GPT-3.5-turbo	0.18	0.16	1	0.16
GPT-4	0.17	0.17	0.16	1

Table 12. Cohen's Kappa with HTML

Model	PaLM 2	Vicuna	GPT-3.5-turbo	GPT-4
PaLM 2	1	0.97	0.77	0.81
Vicuna	0.97	1	0.78	0.81
GPT-3.5-turbo	0.77	0.78	1	0.77
GPT-4	0.81	0.81	0.77	1

Table 13. Relaxed Cohen's Kappa with HTML

Model	PaLM 2	Vicuna	GPT-3.5-turbo	GPT-4
PaLM 2	1	0.064	0.16	0.14
Vicuna	0.064	1	0.0047	0.011
GPT-3.5-turbo	0.16	0.0047	1	0.14
GPT-4	0.14	0.011	0.14	1

Table 14. Cohen's Kappa no HTML

Model	PaLM 2	Vicuna	GPT-3.5-turbo	GPT-4
PaLM 2	1	0.92	0.77	0.74
Vicuna	0.92	1	0.83	0.79
GPT-3.5-turbo	0.77	0.83	1	0.73
GPT-4	0.74	0.68	0.73	1

Table 15. Relaxed Cohen's Kappa no HTML

5 DISCUSSION

5.1 Reflection on Methodological Challenges

Our findings must be interpreted in light of the challenges encountered during the data collection phase. The initial removal of HTML tags, especially those using MathJax for mathematical notation, underscored the importance of data fidelity in studies leveraging LLMs. Differences observed between the evaluations of datasets with and without HTML tags highlight the nuanced understanding LLMs can derive from formatted content. These insights will be invaluable for future work aiming to create enhanced teacher-authored explanations using LLMs.

5.2 HTML Formatting

This study did not initially set out to explore the influence of HTML formatting on the evaluation of teacher-authored mathematical explanations by language models. However, an unintended removal of HTML from the explanations and problem bodies brought this influence to the forefront. Recognizing this alteration, evaluations were rerun using inputs that retained the originally intended HTML formatting. This provided a unique opportunity to examine how HTML formatting impacted evaluations by four state-of-the-art language models: GPT-4, GPT-3.5-turbo, PaLM 2 and Vicuna.

5.3 Evaluation of Teacher-Authored Explanations

For GPT-4, evaluations with HTML formatting consistently scored higher across all categories of the rubric: Accuracy, Explanation Clarity, Mathematical Language and Notation, and Problem Solving Strategy. The statistically significant differences between explanations with and without HTML formatting suggest that GPT-4's evaluations are notably affected by the structural and presentational nuances introduced by HTML. This might indicate that the added structure from HTML provides GPT-4 with added clarity or context. Another perspective could be that GPT-4's training data has made it more sensitive to content with HTML formatting, potentially associating it with higher quality or more structured explanations.

On the other hand, GPT-3.5-turbo showed a varied response to the presence of HTML. Except for the Explanation Clarity category, where a borderline statistically significant difference was observed in favor of HTML-formatted explanations, the model showed no significant difference in evaluations for the other categories.

The results for PaLM 2 and Vicuna further shed light on the variability in model responses to HTML formatting. While PaLM 2's evaluations do not display significant variations between HTML-formatted and non-formatted content across all categories, Vicuna shows distinct patterns in its evaluations, similar to GPT-4, with significant differences emerging due to the presence of HTML.

The differential behaviors among these models, even those from similar lineage like GPT-4 and GPT-3.5-turbo, highlight that subsequent versions or different models may respond differently to nuances in input data. These findings emphasize the need to rigorously test and understand the outputs of these models, especially when subtle changes in input formatting can lead to measurable variations in evaluations.

While the primary objective of this study was not to investigate the influence of HTML, the serendipitous discovery of its removal and the subsequent re-evaluation provided valuable insights into the behavior of several state-of-the-art language models. Future research could further investigate the reasons behind these behaviors and explore the impact of other formatting or structural changes on model evaluations.

5.4 Inter-Rater Reliability

We sought to explore the inter-rater reliability of LLMs to discover if LLMs could replace humans in rating texts. The Cohen's Kappa were generally quite low, between 0 and 0.2 except for the inter-rater reliability between PaLM 2 and Vicuna which had a remarkably high inter-rater reliability. They gave the same rating for 83% of the scores across the 5 categories which HTML text was included. This was only the case in the HTML ratings, where most of the ratings were 3 or 4 which caused the inter-rater reliability to be high. Notably, GPT-3.5-turbo and GPT-4 were did not demonstrate a high inter-rater reliability despite both being in the same family of LLMs.

We believe that adjacent scores often possess a degree of arbitrariness, wherein a rating of 3 might frequently be interchangeable with a rating of 4, and vice versa. Accordingly, we believe that the relaxed Cohen's Kappa metric may provide a more insightful and nuanced assessment of the inter-model agreement in this context. Our analysis using the relaxed Cohen's Kappa demonstrates a high level of agreement between the models, indicating a consistent alignment in their assessment of answer quality, while not always precisely matching the exact score. All relaxed Cohen's Kappa scores were above .67 and many were above .8 indicating a solid inter-rater reliability.

5.5 Future Work

Future studies could expand upon this research in several ways:

- Incorporate human raters for validation to compare their assessments with those of LLMs. We operate under the assumption that if all four LLMs record a higher score for the improved explanations then they are likely improved, this could be confirmed by human raters. We also operate under the assumption that the Inter-Rater reliability of LLMs with other LLMs would be similar to the Inter-Rater reliability of those LLMs with human. We plan to confirm this by having human raters review the explanations and finding the Inter-Rater reliability between each LLM and each human rater.
- We also only evaluated teacher authored explanations, which are generally good and deserve a score of 3 or 4. We plan to extend this approach to evaluate student-authored explanations of varying degrees of correctness.
- Now that we have shown LLM feedback to be at least moderately reliable, we intend to attempt fine-tuning LLMs with Reinforcement Learning with LLM feedback and evaluate the result.

5.6 Limitations

Many of the teacher-authored explanations should likely be scored fairly high, at perhaps 3 or 4 in each category. While there are exceptions, such as explanations which were wrong, we generally expect scores to range higher rather than lower.

Our study was not without its challenges. The unintentional stripping of HTML tags from our initial dataset led us to redefine the scope of our research. While we believe our findings provide valuable insights into the evaluation of teacher-authored explanations by LLMs, the limited exploration of LLM-generated improvements, due to the mentioned challenges, is an area we aim to address in future research.

6 CONCLUSIONS

The unintentional omission of HTML from teacher-authored mathematical explanations provided a unique insight into the evaluation patterns of language models on structured content. GPT-4 showed a consistent inclination for HTML-formatted explanations across all evaluation criteria. Conversely, GPT-3.5-turbo exhibited a more mixed response, with a discernible preference for HTML in the Explanation Clarity category. PaLM 2 displayed negligible variations between formatted and unformatted content. Vicuna, interestingly, demonstrated significant deviations, with a particular aversion to HTML in the Accuracy and Explanation Clarity categories, but a strong affinity in Mathematical Language & Notation.

These disparate responses accentuate the necessity to comprehend the idiosyncratic behaviors of different models. Even nuanced changes in input presentation can instigate distinct evaluation patterns. The inadvertent discovery of HTML's influence in this study serves as a gentle reminder of the broader lesson: the imperative to rigorously test and interpret AI models, especially given their inherent sensitivities to content intricacies.

While each LLM exhibited commendable performance in terms of IRR as shown by the relaxed Cohen's Kappa, it is clear that each model has distinct strengths and areas for improvement. Hopefully this study can help guide other educational researchers in selecting the most appropriate LLM for their specific needs. For tasks requiring precise mathematical language without the overhead of HTML, Vicuna stands out. Meanwhile, PaLM 2 offers a balanced performance across various criteria. It's essential for users looking to adopt LLMs to weigh these considerations when choosing to employ these models in educational contexts.

REFERENCES

[1] Heidi Goodrich Andrade. 2000. Using rubrics to promote thinking and learning. *Educational leadership* 57, 5 (2000), 13–19.

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.

[3] Benjamin S Bloom, Max D Engelhart, EJ Furst, Walker H Hill, and David R Krathwohl. 1956. *Handbook I: cognitive domain.* David McKay, New York.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[5] David F Feldon, Briana Crotwell Timmerman, Kirk A Stowe, and Richard Showman. 2010. Translating expertise into effective instruction: The impacts of cognitive task analysis (CTA) on lab report quality and student retention in the biological sciences. *Journal of research in science teaching* 47, 10 (2010), 1165–1185.

[6] John Hattie. 2008. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* routledge, New York.

[7] Slava Kalyuga. 2007. Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review* 19 (2007), 509–539.

[8] Andreas Lachner and Matthias Nückles. 2016. Tell me why! Content knowledge predicts process-orientation of math researchers' and math teachers' explanations. *Instructional Science* 44 (2016), 221–242.

[9] R. J. Light. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin* 76, 5 (1971), 365–377. https://doi.org/10.1037/h0031643

[10] Deborah Loewenberg Ball, Mark Hoover Thames, and Geoffrey Phelps. 2008. Content knowledge for teaching: What makes it special? *Journal of teacher education* 59, 5 (2008), 389–407.

[11] Richard E Mayer. 2004. Should there be a three-strikes rule against pure discovery learning? *American psychologist* 59, 1 (2004), 14–19.

[12] M David Merrill. 2002. First principles of instruction. *Educational technology research and development* 50 (2002), 43–59.

[13] Mathias Norqvist. 2018. The effect of explanations on mathematical reasoning tasks. *International Journal of Mathematical Education in Science and Technology* 49, 1 (2018), 15–30.

[14] Jean Piaget. 1977. *The development of thought: Equilibration of cognitive structures.(Trans A. Rosin).* Viking, Oxford, England.

[15] Ethan Prihar, Morgan Lee, Mia Hopman, Adam Tauman Kalai, Sofia Vempala, Allison Wang, Gabriel Wickline, Aly Murray, and Neil Heffernan. 2023. Comparing different approaches to generating mathematics explanations using large language models. In *International Conference on Artificial Intelligence in Education*. Springer, Springer Nature, Switzerland, 290–295.

[16] Janine T Remillard. 2005. Examining key concepts in research on teachers' use of mathematics curricula. *Review of educational research* 75, 2 (2005), 211–246.

[17] Alexander Renkl. 2014. Toward an instructionally oriented theory of example-based learning. *Cognitive science* 38, 1 (2014), 1–37.

[18] Lindsey E Richland, James W Stigler, and Keith J Holyoak. 2012. Teaching the conceptual structure of mathematics. *Educational Psychologist* 47, 3 (2012), 189–203.

[19] Barbara Rogoff. 1990. *Apprenticeship in thinking: Cognitive development in social context.* Oxford University Press, New York.

[20] James H. Stronge. 2018. *Qualities of effective teachers / James H. Stronge.* (third edition. ed.). ASCD, Alexandria, VA USA.

[21] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes.* Harvard University Press, Cambridge, MA.

[22] Jörg Wittwer and Alexander Renkl. 2008. Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist* 43, 1 (2008), 49–64.