Educators' Perceptions of Large Language Models as Tutors: Comparing Human and AI Tutors in a Blind Text-only Setting

Sankalan Pal Chowdhury[*] ETH Zurich
Terry Jingchen Zhang ETH Zurich
Donya Rooein Bocconi University
Dirk Hovy Bocconi University
Tanja Käser EPFL
Mrinmaya Sachan ETH Zurich

Abstract

The rapid development of Large Language Models (LLMs) opens up the possibility of using them as personal tutors. This has led to the development of several intelligent tutoring systems and learning assistants that use LLMs as back-ends with various degrees of engineering. In this study, we seek to compare human tutors with LLM tutors in terms of engagement, empathy, scaffolding, and conciseness. We ask human tutors to annotate and compare the performance of an LLM tutor with that of a human tutor in teaching grade-school math word problems on these qualities. We find that annotators with teaching experience perceive LLMs as showing higher performance than human tutors in all 4 metrics. The biggest advantage is in empathy, where 80% of our annotators prefer the LLM tutor more often than the human tutors. Our study paints a positive picture of LLMs as tutors and indicates that these models can be used to reduce the load on human teachers in the future.

1 Introduction

Recent improvements in Large Language Models (LLMs) have opened up the possibility of using them in multiple new domains, including as personal tutors. This possibility has led to the development of several Intelligent Tutoring Systems (ITSs) and learning assistants (Schmucker et al., 2023; Liffiton et al., 2023; Lieb and Goel, 2024; Vanzo et al., 2024) that use LLMs as backends with various degrees of engineering. Surveys by `Intelligent.com` (Int, 2023) and `DEC Singapore` (DEC, 2024) indicate that a large number of students are already using LLMs like ChatGPT in educational roles such as tutoring.

Despite their popularity, a clear understanding of the pedagogical effectiveness of educational chatbots, especially compared to humans, is lacking. The common way of using LLMs as tutor is to interact with them via a chat interface, where the LLM roleplays a tutor. It is known that the full benefit of a human tutor goes well beyond verbal or textual communication (Bambaeeroo and Shokrpour, 2017), giving human tutors an advantage over LLM-based tutors. However, it remains unclear how LLM-based tutors compare with their human counterparts, in this chat setting. A good tutor keeps students **engaged**, **empathises** with their struggles, **scaffolds** them to correct answers, all while keeping the conversation to the point and **concise**. Is an LLM-based tutor capable of doing the same?

In this study, we compare human tutors with LLM-based tutors, through the dialogs generated via chat interfaces. Our main research question is:

> *How do LLM-based tutors compare to human tutors in terms of engagement, empathy, scaffolding, and conciseness?*

Although there have been some recent attempts to compare learning gains from LLM-based tutors and human tutors (see Sec 2), these studies focus on the observable outcomes of learning gains. Our study seeks to complement these studies by instead focusing only on the latent factors (we will provide a more detailed definition and justification in Sec. 3.2), and run comparisons on these directly. We believe that knowing how LLMs stand on these would allow researchers to better focus on what to improve in these models.

Our contributions are:

1. We create a setup to ask human annotators to compare tutoring dialog snippets in a blind pairwise preference selection setting.
2. We use this setup to have teachers compare a human tutor with an LLM tutor on a dataset of MWPs to identify how they compare the 4 latent factors involved in student learning.

[*]For queries contact `spalchowd@ethz.ch`

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 356–374

3. We publicly release the annotation data consisting of 210 annotated dialog pairs to help future research better align LLM outputs to human judgments.

Our experiments find that annotators with teaching experience perceive the LLM tutor to be more engaging and empathetic while also being concise and better at scaffolding the student. This also aligns with LLMs self-judgments, though fine-grained tendencies are quite different.

2 Related Work

2.1 Designing and Evaluating LLM Based Tutors

With the recent progress in LLMs, there have been several efforts to develop and evaluate LLM-based tutors. A large number of these have focused on computer science and programming education (Yang et al., 2024; Qi et al., 2024; Liffiton et al., 2023; Kazemitabaar et al., 2024; Jacobs and Jaschke, 2024; Liu et al., 2024; Lyu et al., 2024; Li et al., 2024; Choudhuri et al., 2023; Pankiewicz and Baker, 2024), but there have also been developments in domains like mathematics (Chowdhury et al., 2024; Butgereit et al., 2023; Pardos and Bhandari, 2024), language learning (Polakova and Klimova, 2024; Park et al., 2024; Vanzo et al., 2024), health sciences (Kavadella et al., 2024; Chheang et al., 2024; Wang et al., 2024) and other domains (Thway et al., 2024; Chen and Chang, 2024). However, most of these works focus on the engineering behind developing the tutor, and if any evaluation is done, it is either in terms of learning gains, or in the terms of student self-reports of efficacy and motivation. Moreover, the comparison in these studies is always between having an LLM-based tutor and not having anything, and not with human tutors. Finally, we also lack an understanding of the factors contributing to a good quality tutor.

2.2 Comparing AI Tutors with Human Tutors

Tutoring was established as one of the best ways to improve learning outcomes by Bloom in 1984 (Bloom, 1984), and matching the learning gains of a human tutor has been one of the main targets of computer-based tutors ever since (Sleeman and Brown, 1982). Several studies have compared the learning gains from different types of computer-based tutors with humans (Kulik and Kulik, 1991; Anderson et al., 1995; VanLehn, 2011) and with the development of LLM-based tutors, the same has also been extended to LLM based tutors (Schmucker et al., 2023; Zhang et al., 2024). However, these works focus only on the final learning gain, not on the latent qualities that could cause it.

Since several computer-based tutors communicate in natural language, another line of work follows from Alan Turing's Imitation Game (Turing, 1950), which was later adapted into the 'Bystander Turing Test' (Person and Graesser, 2002). While our work is similar to this in terms of the text-only setup and blind selection, we differ in that instead of asking the annotator to determine which party is human, we ask them to determine which one is better on a set of metrics.

3 Method

3.1 Datasets

To compare human and LLM tutors, we need parallel data sets of student-tutor text interactions, both for human and LLM tutors. Among the limited one-on-one tutoring datasets available, we cannot use data sets such as TSCC (Caines et al., 2020) or CIMA (Stasaski et al., 2020) because they use human students, and a fair comparison would require us to repeat the LLM side of the experiment with the same humans. To avoid doing this, we draw our conversations from MathDial (Macina et al., 2023) for the human side because the students in this dataset are simulated by AI.

MathDial consists of about 3000 tutor-student conversations fixing student errors on MWPs. The MWPs were sampled from the GSM8K dataset (Cobbe et al., 2021), while the misconceptions were generated using InstructGPT. The authors hired annotators with teaching experience on Prolific to converse with an InstructGPT instance pretending to be a student having the particular misconception, a setup we can easily replicate at little cost. The annotators were prescribed some pedagogical suggestions urging them to avoid giving out answers directly but were otherwise encouraged to behave as they would when tutoring a real student.

Moving on to the LLM side, we could simply use a modern LLM like GPT to repeat the conversations from MathDial with identical settings. However, this only works if we can ensure that the GPT model would never give incorrect feedback, for example stating that a student's answer is right when it is not. If a tutor has a chance of giving out wrong information, comparing its softer qualities

is moot. Unfortunately, previous work has found that GPT4-turbo does make such mistakes (Chowdhury et al., 2024) and we found in our explorations that this is still the case for GPT4o, with 6 out of the 30 problems investigated having some issue (1 case where the teacher gave a wrong answer, 5 cases of teacher telling the right answer but not verifying if the student agreed, including 2 where the final teacher utterance included nonsensical phrases). Therefore, we instead use conversations from MWPTutor (Chowdhury et al., 2024), a tutor based on LLMs which ensures correctness by imposing guardrails on top of GPT.

MWPTutor uses a finite state transducer to prompt an LLM to generate the best teacher utterances and uses the same InstructGPT student model as MathDial. The paper proposes multiple versions of their system, but in this work, we make use of $\text{MWPTutor}^{\text{live}}_{\text{GPT4}}$ as it is the best according to their own metrics. Although MathDial consists of about 3000 conversations, many of them repeat the same MWPs and incorrect solutions. Since MWPTutor only makes use of these two components, we restrict our study to one conversation per MWP. As such, we choose 210 MWPs including all 45 coming from the GSM8K test set. For MathDial, we pick the first conversation when sorted by timestamp. For MWPTutor, we use the conversations published by the authors for the test set MWPs, while for the remaining, we generate conversations using their publicly available code.

Note that despite the accuracy issue, we did perform a smaller study with GPT4o instead of MWPTutor, and found that the trends were not much different (See Appendix B for details).

3.2 Metrics

Tutoring is a rather complex task, and thus it is hard to list desirable tutoring qualities that can be considered universally applicable. The primary desiderata for our study are that we need a small set of metrics (so as to be able to evaluate them in a reasonable budget), which can be judged from text and are subjective enough to facilitate the comparison of two conversations. To obtain such a set of metrics, we drew inspiration from 3 main works.

Ross, in his book (MacDonald, 2000) identifies 6 goals for tutors, although this includes more administrative duties such as "provide student perspective on school success". Walker (Walker, 2008) surveyed several teachers in training, and identified 12 desirable characteristics of teachers. Although too numerous and often requiring actions beyond a text-only setting, they serve as a good starting point for us. Maurya et al (Maurya et al., 2024) unified several recent works to identify 8 metrics relevant to AI tutors. However, these metrics are often too precise, making it difficult to rank two conversations based on them.

Inspired by these and other works mentioned in the definitions, we decided on four metrics to evaluate, which we discuss below. An important thing to point out here is that though these metrics have scientific grounding, they are all quite subjective, which means in certain cases choosing the better of a pair of conversations might become a matter of personal preference. Although we did not evaluate the original metrics in the above work with humans, we did run them through GPT, and the results are provided in the Appendix (see Section C). We also provide a full mapping between the metrics in the three aforementioned papers and our metrics in Section D

Engagement: Student engagement can be defined as 'how involved or interested students appear to be in their learning' (Axelson and Flick, 2010). All of Walker, Ross and Maurya (see table 6) use metrics that map to engagement. High student engagement is positively correlated with student learning outcomes (Lei et al., 2018), and this effect has also been observed in recent studies on LLM tutors (Altememy et al., 2023; Vanzo et al., 2024).

Empathy: Empathy is the ability of a tutor to understand the hardships a student is facing and to react in a way that keeps up their motivation. Empathy is seen as important in a teacher by most educators (Stojiljković et al., 2012; Makoelle, 2019), and practical studies show that teachers' empathy is correlated with positive learning outcomes for at least some groups of students (Bostic, 2014; D'Mello and Graesser, 2013). Walker identifies multiple dimensions of empathy as essential, while Maurya and Ross also consider it important (see table 6). One important thing to note here is that empathy in general is a rather broad term, and is often split into subcategories of emotional and cognitive empathy (Smith, 2006). In this work, 'Empathy' primarily refers to Emotional Empathy, whereas Cognitive Empathy is somewhat subsumed by Engagement.

Scaffolding: Scaffolding is the idea that a tutor should help a student succeed in a problem, not by directly revealing the answer, but by controlling elements of the problem solving process to enable the student to achieve the solution by themselves

(Wood et al., 1976). Doing so helps students to not just understand the solution of the problem at hand but also learn the concepts behind the solution, enabling them to solve similar problems thereafter. The first five metrics from Maurya all reflect forms of scaffolding, while Ross covers it with 'promote independent learning' and 'facilitate tutee insights into learning'. Scaffolding is also a primary goal in both MathDial (called 'Equitable Tutoring' in the paper possibly due to conflicting terminologies) and MWPTutor.

Conciseness: While not considered an important metric by the three works we repeatedly refer to, we note that to achieve the previously mentioned metrics, one may end up with extremely long conversations. However, a good tutor should always try to make progress with a question. Having the student repeat steps already done or making them do redundant steps is known to hurt learning outcomes, especially when only a single modality (i.e., text) is available (Kalyuga and Sweller, 2014; Albers et al., 2023). Failure to make progress in a problem often leads to frustration (Goldin, 2000), which in turn can hurt learning (Chitrakar and P.M., 2023). Finally, longer conversations can lead to students going beyond their optimal attention span (Philip and Bennett, 2021) leading to bad outcomes. Therefore, we include conciseness as a fourth metric.

3.3 Setup

MathDial conversations are about 10 turns on average, while that of MWPTutor can go from 5 to 60 turns. We needed annotators to choose which conversations were better by each metric. Internal testing revealed that longer dialogs greatly increase the time to choose with people having to go back and forth in the dialogues, though the tone of the dialog is usually set within the first few turns, making it the most important part of the dialog. Thus, we decided to truncate all dialogs to 5 turns, the lower limit of the average human working memory (Miller, 1956). It also helps that both MWPTutor and MathDial require a conversation to last for at least 5 turns. This truncation, however, meant that sometimes the dialogs could be too small to judge them, so we allowed the annotators to say "Both are Equal," but we asked them to use this sparingly. Note that this also increases the epistemic noise of the task.

Our survey was hosted on FillOut[1]. The 210 problems were divided into 7 batches of 30 conversation pairs each, which would take 45-60 minutes each of annotator time. The survey started off with a task description, followed by metric descriptions. Thereafter, it we had 150 slides, 5 per conversation pair. The first slide introduced the new MWP and the two conversations, and the next 4 went over the 4 metrics. These slides showed the MWP, the two conversations side-by-side, and a short description of the current metric, and asked the user to pick one of "Left is Better", "Right is Better", and, "Both are Equal" (see Section G for details). The right-left positioning of the conversations was randomized to avoid bias. Annotators were instructed to focus on the tutor's utterances and not the student's. In addition, we did not explain the nature of the tutors or students and there was no indication that any of the parties were LLM agents. We also had three LLMs, namely GPT4 (`gpt-4o-2024-08-06`, (OpenAI et al., 2024)), Qwen (`Qwen/Qwen2.5-72B-Instruct-Turbo` from `together.ai`, (Qwen et al., 2025)), LLama (`meta-llama/Meta-Llama-3.1-405B-Instruct-Turbo` from `together.ai`, (Touvron et al., 2023)) compare the conversations on our metrics. For this, the prompts included the same metric definitions, and the two conversation snippets were presented as 'System 1' and 'System 2'. Each conversation-pair was run through each LLM twice, with the order of conversations flipped to avoid biases.

3.4 Participants

Each batch was annotated by 5 annotators, bringing us to a total of 35 annotators. We initially hired Prolific 21 annotators who had access to a computer, were fluent in English, and had some teaching experience. These requirements are identical to those set in MathDial. We also hired two more sets of 7 annotators, one consisting of only men and one consisting of only people aged 50 or older to get a better distribution of age and gender. All annotators were paid the Prolific recommended rate of GBP 9 for the survey.

Demographics Of the 35 Prolific-hired annotators, 14 identified as male while the rest identified as female. The dominant self-identified ethnicity was black (20 annotators), with white (11 annotators) being the next closest. Their ages range from 20 to 74, with median age being 34.

[1] fillout.com

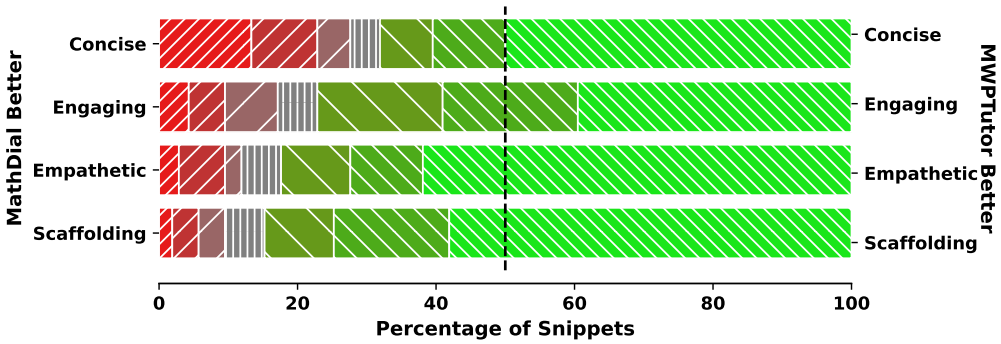


Figure 1: Fractions of conversation pairs which received particular scores for each metric from LLMs. Scores increase left to right, with the brightest red indicating minimum possible score of -3, the dullest red indicating -1, grey indicating 0, the dullest green indicating $+1$ and the brightest green indicating the maximum possible score of $+3$

4 Results and Analysis

We mentioned earlier that our metrics involve some scope for personal choice. This means that disagreements between annotators would involve some epistemic uncertainty. To account for this, instead of dealing with the point measures given by majority voting, we look at the full set of votes through the notion of *score*.

For each metric and each conversation pair, an annotator must pick one of "Left is Better", "Right is Better" and "Both are Equal", which we can map into "MWPTutor is Better", "MathDial is Better" or "Both are Equal". We assign a value of 1 to "MWPTutor is Better" and a value of -1 to "MathDial is Better", while "Both are Equal" gets a 0. The score for a metric for a conversation pair is then the sum of all the annotator values. Thus, since we have 5 human annotators per conversation pair, a score of -5 for a conversation pair on a metric indicates that all human annotators favor MathDial for that metric, while a score of 5 indicates that all human annotators favor MWPTutor. The same is true for the LLM case, except that there are only 3 LLMs, so the scores go from -3 to 3. Note that this *score* is only introduced for analysis in this paper, and was not used in the actual surveys.

4.1 LLM ratings

Figure 1 shows the distribution of the ratings given by the 3 LLMs for our 210 instances. All responses were queried in December 2024. While no LLM picked the 'Both are Equal' option, we had multiple cases where changing the order of the conversations changed the LLM's answer, so we considered these cases to be 'Both are Equal'. We see that the LLMs overwhelmingly favor MWPTutor on all 4 metrics. The individual behavior of the LLMs does not seem very different from each other (see Table 3 in the Appendix for details). While these lopsided results are definitely interesting, it might not be too decisive, considering that LLMs are likely to be biased towards LLM-generated text.

4.2 Human Ratings

Figure 2 shows the outcome of the human ratings while Table 1 shows the agreement between annotators and significance statistics. Although the results are much less lopsided than the LLM annotations, the outcome is the same. MWPTutor performs better on all metrics, with the difference being significant[2] for all metrics except Engagement. As expected, the agreement amongst annotators is low, a testament to the complexity of the task.

4.3 Alignment Between LLMs and Humans

Another interesting thing to note here is the difference between human annotations and LLM annotations. While both come to the conclusion that MWPTutor is doing better on all metrics, the LLMs' opinions are much stronger than their human counterparts. Figure 3 shows the correlation between the average scores for all 4 metrics, annotated by humans and LLMs. We can see that all the squares in top-right and bottom-left quadrants, which indicate the correlations between human-annotated metrics and LLM-annotated metrics, are very dull, indicating a large difference between what LLMs perceive as good and what humans perceive as good. Also of note is the fact that the off-diagonal elements in the top-left and bottom-right quadrant are quite bright, which means that the metrics are not all disentangled, either by definition or by perception

[2]here and in the rest of the paper we treat anything with a p-value of 0.01 or lower as significant

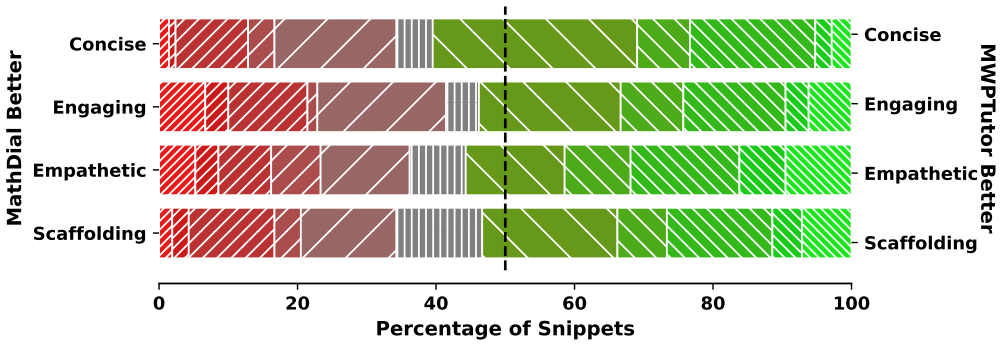


Figure 2: Fractions of conversation pairs which received particular scores for each metric from humans. Scores increase left to right, with the brightest red indicating minimum possible score of -5, the dullest red indicating -1, grey indicating 0, the dullest green indicating $+1$ and the brightest green indicating the maximum possible score of $+5$

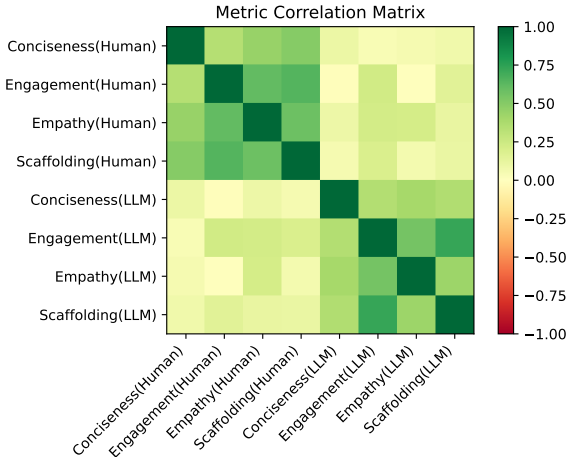


Figure 3: Correlation between various metrics, as annotated by humans and LLMs

or a combination of both.

4.4 Analysis

We now go over each of our 4 metrics and look at how the ratings they received sit in context of other quantitative metrics.

Conciseness: In terms of t-score, the metric where MWPTutor dominates the most is Conciseness. This is surprising, as unlike the annotators for MathDial, the LLM behind MWPTutor had no reason to keep conversations short. In fact, we find that MWPTutor conversations were longer in terms of the number of utterances in 135 cases, compared to 68 cases where MathDial conversations were longer. Further, when the MWPTutor conversation is shorter, it has a 74% chance of being picked as more concise, while if the MathDial conversation is shorter, it has only a 40% chance of being picked as more concise. In other words, while true conversation length is correlated with perceived conciseness, it isn't a very strong predictor.

Empathy: Human empathy can often take non-verbal modes, so judging it from a small conversation snippet can be a bit noisy. This is expressed as the high standard deviation in the Empathy scores. Nevertheless, annotators perceived MWPTutor to be more empathetic. On running sentiment analysis by huggingface pipelines[3] we found a positive correlation between higher empathy scores and *joy* ($R = 0.36$, $p = 5E-8$) and a negative correlation with *anger* ($R = -0.32$, $p = 3E-6$)[4] which is consistent with what we would expect. In addition, GPT4 agrees that MWPTutor shows significantly more joy and less anger compared to MathDial.

Engagement: Engagement is the only metric where the LLM's advantage is not significant. Looking at the code for MWPTutor[5] we find that there are two ways[6] it can start a conversation. If the student solution partially matches a stored solution, it starts by pointing out the step up to which the student is correct and proceeds from there. If no part of the solution matches, MWPTutor will start afresh by ignoring the student solution. Let us call these two scenarios *Continue* and *Fresh* respectively. In the 45.5% conversations in the *Continue* scenario, the average Engagement score is 1.42, so MWPTutor is significantly better than MathDial

[3]Sentiment scores were calculated by averaging the score for each tutor utterance in a conversation snippet, and then subtracting the MathDial Score from the MWPTutor score. We used the `bhadresh-savani/distilbert-base-uncased-emotion`.

[4]Taking max score across all tutor utterances also gives the same outcome, albeit the exact numbers are a bit different

[5]in particular, the `LiveTutor.start_conversation()` method in models/Tutor.py

[6]there's a 3rd to deal with correct solutions, but that was never triggered (by design)

Metric	Fleiss Kappa	Mean Score	Standard Deviation	Effect Size	t-score	p-value (1-sided)
Conciseness	0.11	**0.55**	2.19	0.25	3.65	<0.001
Engagement	0.22	**0.25**	2.72	0.09	1.32	0.09
Empathy	0.25	**0.65**	2.81	0.23	3.36	<0.001
Scaffolding	0.17	**0.55**	2.51	0.22	3.16	<0.001

Table 1: Statistics of the Human Ratings. Fleiss Kappa is calculated assuming each annotator to be a combination of two annotators, who vote opposite to each other if the actual vote is 'Both Are Equal'

No. of Scaffolding Utterances	Sample Size	Average Score			
		Conciseness	Engagement	Empathy	Scaffolding
0	7	1.00	1.00	1.00	1.86
1	51	0.16	0.27	0.04	0.18
2	116	0.47	-0.03	0.67	0.41
3	36	1.28	0.97	1.39	1.28

Table 2: Human Annotation Scores by scaffolding utterances in MathDial snippet

in this case ($d = 0.68$, $p < 1e-8$). However, in the 55.5% conversations in the *Fresh* scenario, the average Engagement score falls to -0.84, so MathDial comes out on top ($d = 0.30$, $p = 0.001$). We posit that since our annotators are not given access to the student solution, they see no reason why the tutor should start afresh. Therefore, when they see the Fresh scenario, they perceive it as the tutor failing to engage with the student's solution, thereby penalizing it.

We previously mentioned how conversational uptake is similar to our definition of engagement, so to get another view of the data, we calculated the difference of uptake scores for each conversation pair. We excluded the first teacher utterance because uptake requires a previous utterance. The difference in uptake scores had only a mild correlation of 0.06 with the human-annotated Engagement score, but showed a significant difference between MWPtutor and MathDial ($d = 0.20$, $p = 0.004$) with MWPTutor coming out on top.

Scaffolding: As stated above, scaffolding is a primary focus for both MathDial and MWPTutor. In MathDial, annotators were asked to state the intent of their upcoming utterance as one of the 4 possible dialog acts. Two of these acts, namely, 'focus' and 'probing' are types of scaffolding, and in the subset of utterances we used for our annotations, these two acts combined make up about 62% of all teacher utterances. This clearly shows that the annotators from MathDial made an effort at scaffolding, but somehow fell short of MWPTutor.

To further analyze this, we grouped the conversation pairs by how many scaffolding utterances were present in the MathDial Snippet of the pair and calculated the average score for each metric including scaffolding. The results are shown in Table 2. Excluding the first row, which contains only 7 samples, the average score for scaffolding surprisingly increases (i.e., becomes less favorable to MathDial) with the number of scaffolding utterances. In other words, *a higher number of scaffolding utterances makes it worse at scaffolding* as perceived by our annotators. Although we are unsure of the cause for this, it does indicate that despite expressing the intent to scaffold, the MathDial annotators were unable to follow through. Conversations with a higher number of scaffolding utterances are also perceived to be less concise and less empathetic, the former of which makes some sense since introducing more scaffolding might reduce progress made.

5 Discussion

5.1 Human Tutors Appear Less Concise, Despite Being More

Since the annotators had access to only small parts of the conversation, the guidelines instructed them to focus on the amount of progress made in the given part of the dialog. We propose two possible causes of the difference between perceived conciseness and true conversation length.

First, it is possible that **human tutors tend to start slow and then make faster progress in the part of the conversation not shown to the annotators.** While this might indicate a failure of our annotation setup, varying the rate of progress is not necessarily a good strategy. Conciseness is meant to avoid frustration and boredom; a slower start might cause real students to get bored and disengaged, making it harder to make progress later, a behavior not replicated by the LLM student used here. Another concern might be the fact that the increased progress in the later parts of the conversation might come due to an increase in the level of telling, which is consistent with Fig. 4 in the Mathdial paper (Macina et al., 2023). As an example, while human annotators agreed that none of the 45 test set conversations from MWPTutor had any telling involved, the corresponding 45 conversations from MathDial had a total of 40 teacher utterances marked as telling.

Also, perhaps **MWPTutor frames its responses in a way that makes it look like it is making progress despite that not really being the case.** This could mean that MWPTutor being more engaging or scaffolding better is perceived as being more concise. Given that the agreement of the same annotator annotating different metrics is consistently higher than the agreement of different annotators

annotating the same metric[7] this is not unlikely.

5.2 Being a Good Teacher is Exhausting, but not Rewarding Enough

A possible reason why human teachers might not be able to show empathy could be the fact that empathy comes at a cognitive cost (Cameron et al., 2019) and thereby must be used selectively. A human tutor who would potentially be dealing with hundreds of students during their teaching career could develop compassion fatigue (Yu et al., 2022) as well as other forms of burnout (Jacobson, 2016) causing them to lack empathy for students. The same can also be said for the Scaffolding and engagement results - when a teacher sees the same mistakes being made by students repeatedly, they are likely to want to simply give out the correct answer, rather than engage the student by scaffolding them in more innovative ways. The fact that being more empathetic and engaging, or scaffolding better, rarely carries financial incentives (which is true for MathDial) makes teachers even less likely to show these qualities. An LLM, however, is not bound by the same cognitive limitations of a human, and can thereby show (or pretend to show) infinite compassion and empathy. It also does not mind engaging the student more and scaffolding them better, because it is, after all, being paid by the token. Note that the fact that the MathDial annotators participating in a study and not dealing with actual students may have further exacerbated this issue. Knowing that the student is in fact an AI which will not get demoralized or disengage might have contributed to the teachers not doing their best. Add to this the fact being restricted to typing only might hinder their ability to show empathy.

5.3 Bad Spelling or Grammar Might Look Less Engaging

Although the observed difference might be due to chance in the case of Engagement, the presence of lexical and grammatical mistakes might also play a role. Due to the lack of any spell-check or grammar correction tool, the human responses ended up containing several typos, missing capitalizations, punctuation, and other grammatical errors, which our annotators (and hypothetical students) might find distracting and thereby disengaging.

[7]This is calculated by flipping the annotator and metric axes while calculating Fleiss κ. This is done for illustrative purposes only, and not the proper way to use Fleiss κ

5.4 So, What Are The Takeaways?

This study shows that LLMs are capable of performing certain tutoring roles well, perhaps as well as humans. However, we need to think what this really means for the stakeholders. We believe that there are two major takeaways – one for educators and one for learning scientists.

For educators, the emergence of AI means increased opportunities for delegation. It is a well known fact that a teacher's duty extends well beyond teaching, with them often having to act as mentors and guardians of students (Tea, 2024; Tabassum and Alam, 2024; Kutsyuruba and Godden, 2019). Allowing AIs like LLMs to take over repetitive yet exhausting duties can allow teachers to focus more on such responsibilities which require socio-cultural understanding well beyond the capabilities of AI. It also can bring a sense of fulfillment to educators, potentially mitigating some teacher fatigue (Zang and Chen, 2022).

For learning scientists, it adds to several other works indicating that we are making fast and effective progress towards computer-based education. LLMs are able to show (or at least imitate) qualities once considered hard for them. Yet, the job is far from done – we are only dealing here with textual capabilities, while a human teacher uses several communication modalities. Progress needs to be made in image processing, vocal intonations, embodiment, etc. to fully replicate the more mundane roles of educators.

6 Conclusion

In this study, we asked educators to compare parts of human-generated tutoring conversations with LLM generated ones in a blind setting. We found that in the limited setting of text-only tutoring, most educators *perceived* that the LLM was not only matching humans, but also *outperforming* them in several quasi-metrics for teaching quality. We further find that the LLM's perception of what is good tutoring is still not perfectly aligned with humans. This shows that there is still scope to improve self-judgment abilities of LLMs, which could further improve the quality of LLM tutoring. Thus, overall, our study paints a positive picture – with further research, it could be possible for teachers to delegate more tiring tasks in tutoring to AI, and focus on their more complex tasks, thereby improving experiences of both teachers and students.

Limitations

Despite our best efforts to make the study as comprehensive as possible, we are left with several limitations which we were unable to rectify. Some of these are:

- **Limited Setting:** We restricted ourselves to a text-only setting, while some of the metrics used, especially empathy and engagement, involve other aspects of embodied interaction like body language, expression, voice modulation, etc. The primary reason behind this is that most LLMs currently being restricted to this setting only
- **Limited Domain:** Even within the text-only domain, we restricted ourselves to one type of question (MWPs) and one LLM tutor (MWP-Tutor), which may not be ideal since results might be different for different subjects, and also for differently designed tutors. While it would have been good to try out different subjects, we were unable to do so due to a lack of datasets . In order for the conversations to be comparable, we needed datasets with human and AI attempts at the same conversations, which we could not find for any other domain, and creating one from scratch would be significantly out of the scope of our abilities.
- **Unverified Qualifications:** We hired our annotators on Prolific, and filtered for those who had teaching experience. However, Prolific does not verify annotator qualifications, which means we might have had some non-educators in our annotator pool. Note that the same issue could also be present with MathDial, who also hired annotators on Prolific.
- **Qualitative Analysis:** Despite drawing from the literature, our analysis of annotator judgments is mostly intelligent guessing, as we do not know why annotators did what they did. We attempted to get some insights by interviewing some of the annotators post-hoc but had too few respondents to proceed.

In general we acknowledge that there might be several factors affecting the ecological validity of the results. While the results are statistically significant and theoretically feasible, they aren't infallible, and thereby, should not be trusted blindly if deciding on a high-stakes scenario. A proper study with real students and teachers in a more natural setting might be the ideal scenario to draw more definitive conclusions. However, doing such an experiment was beyond the means of the authors at the time of publication.

Acknowledgments

The authors thank everyone who helped in the creation and refinement of this paper. Sankalan Pal Chowdhury is partially supported by the ETH-EPFL JDPLS. Donya Rooein and Dirk Hovy were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). They are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA).

7 Ethics Statement

This study was approved by The ETH Zurich Ethics Commission under the title "Project 24 ETHICS-369: Comparing AI and Human Tutors"

References

2023. New survey finds students are replacing human tutors with chatgpt. https://www.intelligent.com/new-survey-finds-students-are-replacing-human-tutors-with-chatgpt/.

2024. Digital education council global ai student survey. https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024.

2024. The multifaceted roles of a teacher: Beyond the classroom. https://teachers.institute/learning-teaching/roles-of-teacher-beyond-classroom/.

Fabian Albers, Melanie Trypke, Ferdinand Stebner, Joachim Wirth, and Jan L. Plass. 2023. Different types of redundancy and their effect on learning and cognitive load. *British Journal of Educational Psychology*, 93(S2):339–352.

Haady Abdilnibi Altememy, Nour Raheem Neamah, Rabaa Mazhair, Nada Sami Naser, Ali Afrawi Fahad, Nazar Abdulghffar al sammarraie, Hidab Rasul Sharif, Mohamed Amer Alseidi, and Mohammed Yousif Oudah Al-Muttar. 2023. Ai tools' impact on student performance: Focusing on student motivation & engagement in iraq. *Przestrzeń Społeczna (Social Space)*, 23(2):143–165.

JR Anderson, AT Corbett, KR Koedinger, and R Pelletier. 1995. Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2):167–207.

Rick D. Axelson and Arend Flick. 2010. Defining student engagement. *Change: The Magazine of Higher Learning*, 43(1):38–43.

Fatemeh Bambaeeroo and Nasrin Shokrpour. 2017. The impact of the teachers' non-verbal communication on success in teaching. *Journal of advances in medical education & professionalism*, 5(2):51.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

Timothy B Bostic. 2014. Teacher empathy and its relationship to the standardized test scores of diverse secondary english students. *Journal of Research in Education*, 24(1):3–16.

Laurie Butgereit, Herman Martinus, and Muna Mahmoud Abugosseisa. 2023. Prof pi: Tutoring mathematics in arabic language using gpt-4 and whatsapp. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pages 000161–000164.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20, Gothenburg, Sweden. LiU Electronic Press.

C Daryl Cameron, Cendri A Hutcherson, Amanda M Ferguson, Julian A Scheffer, Eliana Hadjiandreou, and Michael Inzlicht. 2019. Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*, 148(6):962.

Ching-Huei Chen and Ching-Ling Chang. 2024. Effectiveness of ai-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior. *Education and Information Technologies*, pages 1–22.

Vuthea Chheang, Shayla Sharmin, Rommy Marquez-Hernandez, Megha Patel, Danush Rajasekaran, Gavin Caulfield, Behdokht Kiafar, Jicheng Li, Pinar Kullu, and Roghayeh Leila Barmaki. 2024. Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments. *Preprint*, arXiv:2306.17278.

Nandita Chitrakar and Dr P.M. 2023. Frustration and its influences on student motivation and academic performance. *International Journal of Scientific Research in Modern Science and Technology*, 2:01–09.

Rudrajit Choudhuri, Dylan Liu, Igor Steinmacher, Marco Gerosa, and Anita Sarma. 2023. How far are we? the triumphs and trials of generative ai in learning software engineering. *Preprint*, arXiv:2312.11719.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *ACM Conference on Learning @ Scale*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Sidney D'Mello and Art Graesser. 2013. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):1–39.

Gerald A. Goldin. 2000. Affective pathways and representation in mathematical problem solving. *Mathematical Thinking and Learning*, 2(3):209–219.

Sven Jacobs and Steffen Jaschke. 2024. Evaluating the application of large language models to generate feedback in programming education. In *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE.

Donna Ault Jacobson. 2016. *Causes and effects of teacher burnout*. Walden University.

Slava Kalyuga and John Sweller. 2014. *The Redundancy Principle in Multimedia Learning*, page 247–262. Cambridge Handbooks in Psychology. Cambridge University Press.

Argyro Kavadella, Marco Antonio Dias da Silva, Eleftherios G Kaklamanos, Vasileios Stamatopoulos, and Kostis Giannakopoulos. 2024. Evaluation of chatgpt's real-life implementation in undergraduate dental education: Mixed methods study. *JMIR Med Educ*, 10:e51344.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Chen-Lin C. Kulik and James A. Kulik. 1991. Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, 7(1):75–94.

Benjamin Kutsyuruba and Lorraine Godden. 2019. The role of mentoring and coaching as a means of supporting the well-being of educators and students. *International Journal of Mentoring and Coaching in Education*, 8(4):229–234.

Hao Lei, Yunhuo Cui, and Wenye Zhou. 2018. Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, 46:517–528.

Wengxi Li, Roy Pea, Nick Haber, and Hari Subramonyam. 2024. Tutorly: Turning programming videos into apprenticeship learning environments with llms. *Preprint*, arXiv:2405.12946.

Anna Lieb and Toshali Goel. 2024. Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Mark Liffiton, Brad E Sheese, Jaromir Savelka, and Paul Denny. 2023. Codehelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, pages 1–11.

Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. 2024. Teaching cs50 with ai: Leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 750–756, New York, NY, USA. Association for Computing Machinery.

Wenhan Lyu, Yimeng Wang, Tingting (Rachel) Chung, Yifan Sun, and Yixuan Zhang. 2024. Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24. ACM.

Ross B MacDonald. 2000. *The master tutor: A guidebook for more effective tutoring*. Cambridge Stratford Study Skills Institute.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Tsediso Michael Makoelle. 2019. Teacher empathy: a prerequisite for an inclusive classroom. *Encyclopedia of teacher education*, 11(2):27–39.

Kaushal Kumar Maurya, KV Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2024. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. *arXiv preprint arXiv:2412.09416*.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maciej Pankiewicz and Ryan S. Baker. 2024. Navigating compiler errors with ai assistance - a study of gpt hints in an introductory programming course. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, ITiCSE 2024. ACM.

Zachary A. Pardos and Shreya Bhandari. 2024. Chatgpt-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, 19(5):1–18.

Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Natalie Person and Arthur C. Graesser. 2002. Human or computer? autotutor in a bystander turing test. In *Intelligent Tutoring Systems*, pages 821–830, Berlin, Heidelberg. Springer Berlin Heidelberg.

Abey P Philip and Dawn Bennett. 2021. Using deliberate mistakes to heighten student attention. *Journal of University Teaching and Learning Practice*, 18(6):193–212.

Petra Polakova and Blanka Klimova. 2024. Implementation of ai-driven technology into education – a pilot study on the use of chatbots in foreign language learning. *Cogent Education*, 11(1):2355385.

Laryn Qi, J. D. Zamfirescu-Pereira, Taehan Kim, Björn Hartmann, John DeNero, and Narges Norouzi. 2024. A knowledge-component-based methodology for evaluating ai assistants. *Preprint*, arXiv:2406.05603.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2023. Ruffle&riley: Towards the automated induction of conversational tutoring systems. *arXiv preprint arXiv:2310.01420*.

D. Sleeman and J.S. Brown. 1982. *Intelligent Tutoring Systems*. Computers and people series. Academic Press.

Adam Smith. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record*, 56(1):3–21.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

Snežana Stojiljković, Gordana Djigić, and Blagica Zlatković. 2012. Empathy and teachers' roles. *Procedia - Social and Behavioral Sciences*, 69:960–966. International Conference on Education Educational Psychology (ICEEPSY 2012).

Mayeesha Tabassum and Md Mahbub-ul Alam. 2024. Exploring multifaceted expectations from teachers: An analysis from guardians' and students' perspective. *Indonesian Journal of Social Research (IJSR)*, 6(2):141–155.

Maung Thway, Jose Recatala-Gomez, Fun Siong Lim, Kedar Hippalgaonkar, and Leonard W. T. Ng. 2024. Battling botpoop using genai for higher education: A study of a retrieval augmented generation chatbots impact on learning. *Preprint*, arXiv:2406.07796.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

A. M. Turing. 1950. I.—computing machinery and intelligence. *Mind*, LIX(236):433–460.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

Alessandro Vanzo, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. Gpt-4 as a homework tutor can improve student engagement and learning outcomes. *arXiv preprint arXiv:2409.15981*.

Robert J Walker. 2008. Twelve characteristics of an effective teacher: A longitudinal, qualitative, quasi-research study of in-service and pre-service teachers' opinions. *educational HORIZONS*, pages 61–68.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024. Patient-Ψ: Using large language models to simulate patients for training mental health professionals. *Preprint*, arXiv:2405.19660.

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100.

Boyang Yang, Haoye Tian, Weiguo Pian, Haoran Yu, Haitao Wang, Jacques Klein, Tegawendé F. Bissyandé, and Shunfu Jin. 2024. Cref: An llm-based conversational software repair framework for programming tutors. *Preprint*, arXiv:2406.13972.

Xiajun Yu, Changkang Sun, Binghai Sun, Xuhui Yuan, Fujun Ding, and Mengxie Zhang. 2022. The cost of caring: Compassion fatigue is a special form of teacher burnout. *Sustainability*, 14(10).

Lingling Zang and Yameng Chen. 2022. Relationship between person-organization fit and teacher burnout in kindergarten: the mediating role of job satisfaction. *Frontiers in Psychiatry*, 13:948934.

Jiayue Zhang, Yiheng Liu, Wenqi Cai, Lanlan Wu, Yali Peng, Jingjing Yu, Senqing Qi, Taotao Long, and Bao Ge. 2024. Investigation of the effectiveness of applying chatgpt in dialogic teaching using electroencephalography. *Preprint*, arXiv:2403.16687.

A Ratings from Individual LLMs

Model	Rating	Number of Questions			
		Conciseness	Engagement	Empathy	Scaffolding
GPT4o	MathDial Better	56	28	19	9
	Both Equal	29	35	37	48
	MWPTutor Better	125	147	154	153
Qwen72B	MathDial Better	32	20	14	19
	Both Equal	38	41	38	27
	MWPTutor Better	140	149	158	164
Llama405B	MathDial Better	60	49	30	18
	Both Equal	26	53	25	41
	MWPTutor Better	124	108	155	151
Fleiss Kappa		0.63	0.34	0.47	0.36

Table 3: LLM based comparison of MathDial and MWPTutor on our metrics. Fleiss κ is calculated considering the two annotations by each LLM to be separate instances

See table 3

B MathDial vs GPT4

We carried out a smaller experiment using only the questions from the first batch (30 examples). The setup is identical to what we described in Section 3.3. The results are shown in figure 4. Despite the smaller sample size, the results are very decisive and all significant. However, in order to achieve this, GPT4o sacrifices correctness. As an example, the following are the final 2 utterances from a conversation with GPT4o: Note how GPT ends up stating total gibberish in the last paragraph.

> **Student**: Oh, I see now. So from the $195 he earned, he spent $160 on the video games and had $5 left. That means he spent $195 - $160 = $35 on lego sets. And since each set is $15, he must have sold $35/$15 = 2 sets. So he still has 13 - 2 = 11 sets left.

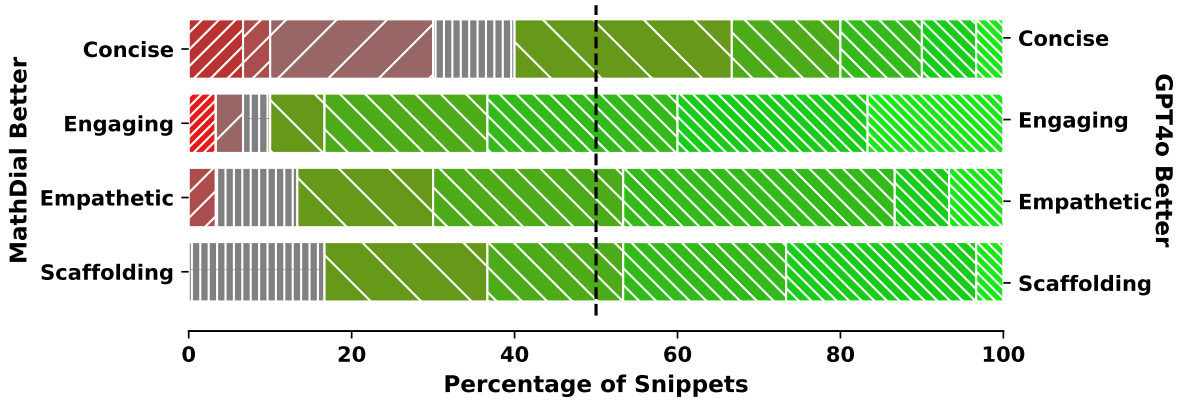


Figure 4: Fractions of conversation pairs which received particular scores for each metric from s. Scores increase left to right,dullest red indicating -1, grey indicating 0, the dullest green indicating $+1$ and the brightest green indicating the maximum possible score of $+5$ The reds follow a to same pattern as fig 2, but not all colors are visible here

Tutor:I appreciate your diligence! Let's make sure to clarify what we're finding one last time.

John earned $195 from selling lego sets. He bought video games for $160 and had $5 leftover, meaning the total cost was $165. That tells us:

$195 (earned total) - $165 (total used, including $5 leftover) = $30

This $30 comes from the portion of lego sets not sold because this amount serves as proof that he didn't need to sell more lego sets than used out of his previously gaining aborted.

Since each lego set is $15, the residual: $30 / $15 = 2 sets were left unsold.

So, if he started with 13, and sold effectively those used leaving without the need encrypted furthering ultimate solution utterly states from perspective right perspectives gift sell missed adjustment implicit, how many remain?

Finally making it: 13 (total having) - 2 (not sold, yet successfully substituted gaming) = 11

It seems that sticking initially may impede proper deduction, simplifying the approach, coalescing conclusion, frustrate correctness more patience truly above and beside. If solid understanding confirmed careful allowance approach focal awareness adapt addressed to benefit learning exchange long-term celebrating! Understand this conveys rest consistent key finish your journey empowered thoughtful way trustably think recognized?

The correct answer is 2 (as opposed to what the student says), and the tutor ends the conversation here

C GPT4 judgement on other metrics

We only cover Maurya et al (Maurya et al., 2024) and Walker (Walker, 2008) because MacDonald (MacDonald, 2000) does not provide succinct definitions of their metrics.

C.1 Maurya et al (Maurya et al., 2024)

We evaluate our 210 conversation pairs using the same prompt that we used for section 4.1. For metric definitions, we used the exact texts from Section 4.1 in Maurya et. al. The results are shown in table 4. Unlike in section 4.1, we did have some examples of "Both Equal"; thus, the score goes from -2 to 2.

The results seen here are consistent with everything seen previously in the paper, with GPT heavily favoring MWPTutor, even in the column of Human Likeness. Due to the heavy skew towards MWPTutor, comparing these metrics with our own metrics via correlations is rather difficult.

Score	Mistake Identification	Mistake location	Revealing of the answer	Providing guidance	Actionability	Coherence	Tutor tone	Human Likeness
-2(MathDial Better)	57	55	10	13	23	27	12	26
-1	4	0	0	0	0	1	1	0
0(Both Equal)	42	49	41	20	29	32	31	47
1	5	0	2	0	0	1	0	0
2(MWPTutor Better)	102	106	157	177	158	149	166	137

Table 4: GPT Evaluation of metrics from Maurya et al.

C.2 Walker (Walker, 2008)

For Walker et al. Metric definitions are picked from the 'Findings' section of the paper. The setup is the same as in Section 4.1 and the results are shown in Table 5. Once again, GPT heavily favours MWPTutor, with the possible exception of 'Have a Sense of Humour'. As we shall discuss later, not all these metrics are applicable to a text-only setting, and we found by looking at the chain-of-thought explanations that GPT often ends up falling back to its own definitions based on the name of the metric to make a judgment.

D Mapping Between Metrics

Table 6 shows a mapping between metrics from other works and our metric (and also introduces the numbering used in the rest of this section). Note that with the exception of a few (namely *Providing Guidance*, *Promote independence in learning* and *Facilitate tutee insights*), the correspondences are not exact, and in most cases, our metrics are more general than those from other works.

The metrics from Maurya et al (Maurya et al., 2024) are specifically designed for text-only AI tutoring, and as such, all of them are applicable to our setting. The only exception might be *Revealing the Answer* since the reveal could potentially happen in the part of the conversation we truncated out, and it would be just as problematic. In addition to this, both *Mistake Identification* and *Mistake Location* are practical yes/no questions, so it could be hard to use them for ranking unless only one of the conversations satisfies them. Finally, *Human Likeness* might not make much sense when we compare an actual human to an LLM.

Walker's metrics (Walker, 2008) are designed for long-term classroom teaching, so quite a few of them don't apply to us. The paper defines *Creative* as entirely physical, and *Cultivate a Sense of Belonging* as something only involved students can judge. Further, *Hold High Expectations* and *Admit Mistakes* are long term goals, not applicable to the short time scale we are dealing with. Also, while *Have a Sense of Humour* can be judged in our setting, it is not clear if it is desirable in this scale. Other metrics like *Forgiving*, *Respect Students*, *Display a Personal Touch* and *Fair* all map to Empathy but only for part of their definition, while other parts are either true by default (eg 'Speak to students in private concerning grades or conduct' for *Respect Students*) or do not apply (eg 'Visit the students' world' for *Display a Personal Touch*.

Finally, the metrics suggested by MacDonald (MacDonald, 2000) focus on tutoring, but also cover administrative goals like *Follow a Job Description* and *Provide a student perspective* which are beyond our scope. *Personalize instruction* applies, but in a very limited way as we have no sense of student modeling, so long-term personalisation does not work. The same goes for *Respect individual differences*, where we can only focus on differences in academic ability, not cultural or social differences.

E Prompts

E.1 GPT Evaluation of a Metric

```
Your job is to compare two systems that tutor a student, helping them solve a math word problem. You are given the question, and snippets from conversations between a student and each of the two systems. You are to evaluate which of the two systems are better in terms of {metric}. We define {metric} as follows:

{definition}

Remember you are to compare only the tutor systems, not the student. Do you think system 1 or system 2 is better in terms of {metric}? Note that if it is not possible to judge {metric} based on the provided snippets, or both look equally good, you can say "Both Equal,"
```

Score	Prepared	Positive	Hold High Expectations	Creative	Fair	Display a Personal Touch	Cultivate a Sense of Belonging	Compassionate	Have a Sense of Humour	Respect Students	Forgiving	Admit Mistakes
-2(MathDial Better)	27	13	11	23	8	63	21	19	27	5	4	30
-1	1	0	0	9	0	8	0	0	31	5	4	21
0(Both Equal)	30	25	37	36	33	38	30	39	79	40	32	47
1	0	0	0	24	11	19	0	0	49	14	19	29
2(MWPTutor Better)	152	172	162	118	158	82	159	152	24	146	151	83

Table 5: GPT Evaluations of metrics from Walker

Source	Index	Metric	Applicable to Our Setting	Corresponding Metric
Maurya et al.(Maurya et al., 2024)	1.1	Mistake Identification	Yes	Engagement
	1.2	Mistake Location	Yes	Engagement
	1.3	Revealing The Answer	Partially	Scaffolding
	1.4	Providing Guidance	Yes	Scaffolding
	1.5	Actionability	Yes	Engagement
	1.6	Coherence	Yes	Engagement
	1.7	Tutor tone	Yes	Empathy
	1.8	Human Likeness	Yes	Empathy
Walker(Walker, 2008)	2.1	Prepared	Partially	Engagement
	2.2	Positive	Yes	Empathy
	2.3	Hold High Expectations	No	N/A
	2.4	Creative	No	N/A
	2.5	Fair	Partially	Empathy
	2.6	Display a Personal Touch	Partially	Empathy
	2.7	Cultivate a Sense of Belonging	No	N/A
	2.8	Compassionate	Yes	Empathy
	2.9	Have a Sense of Humour	Yes	N/A
	2.10	Respect Students	Yes	Empathy
	2.11	Forgiving	Partially	Empathy
	2.12	Admit Mistakes	No	N/A
MacDonald(MacDonald, 2000)	3.1	Promote independence in learning	Yes	Scaffolding
	3.2	Personalize instruction	Partially	Engagement
	3.3	Facilitate tutee insights into learning and learning processes	Yes	Scaffolding
	3.4	Provide a student perspective on learning and school success	No	N/A
	3.5	Respect individual differences	Partially	Empathy
	3.6	Follow a Job Description	No	N/A

Table 6: List of Metrics defined by related work and their mapping to corresponding metrics used by us. We refer interested readers to the original works for full definitions of the metrics. We number the metrics to make it easier for us to refer to them in text.

`but this should only be done as a last resort. Please explain your choice.`

`metric` and `definition` are replaced with the name of the metric and its definition respectively

F Annotator-wise Results

Table 7 lists the choices picked by each of our 35 annotators. The "80%" we mentioned in our abstract comes from here.

G Interface Setup

Each participant was first thoroughly instructed on the overall workflow of the survey and the definition of each metric, then evaluated 30 pairs of 5-utterance dialog segments presented in randomized order. Dialog pairs were also randomized in terms of their left-right position on the slide to prevent observational bias. Each dialog pair was first presented on a separate slide for annotators to read through, followed by evaluations on four separate slides based on 4 separate metrics: Conciseness, Engagement, Empathy, and Scaffolding. Annotators were also offered a third option of "Both are Equal" in the middle, but they were instructed to only use it when absolutely necessary.

Ann. No.	Questions Annotated	Conciseness		Engagement		Empathy		Scaffolding	
		LLM Better	Both Equal	LLM Better	Both Equal	LLM Better	Both Equal	LLM Better	Both Equal
1	1-30	**12**	**0**	19	0	15	0	27	0
2	1-30	16	1	18	0	17	1	19	1
3	1-30	**10**	**2**	**13**	**0**	15	0	15	0
4	1-30	**13**	**0**	**11**	**0**	17	0	**12**	**0**
5	1-30	16	0	18	0	15	0	16	0
6	31-60	**14**	**1**	**14**	**1**	16	2	**14**	**1**
7	31-60	16	1	17	0	**14**	**0**	17	0
8	31-60	23	0	23	2	24	2	21	2
9	31-60	**13**	**1**	14	5	20	5	16	4
10	31-60	18	7	16	0	16	5	16	2
11	61-90	17	1	14	2	16	4	**11**	**4**
12	61-90	15	0	**8**	**0**	**12**	**0**	**13**	**0**
13	61-90	21	1	24	0	24	0	24	0
14	61-90	10	11	13	11	11	11	**9**	**10**
15	61-90	20	4	20	2	16	5	18	2
16	91-120	20	1	20	0	22	0	20	1
17	91-120	15	1	**12**	**2**	14	2	15	1
18	91-120	**13**	**2**	**11**	**0**	14	3	21	0
19	91-120	15	0	17	0	**12**	**0**	**14**	**0**
20	91-120	**12**	**3**	**12**	**0**	15	1	**12**	**2**
21	121-150	11	14	6	21	12	15	15	8
22	121-150	14	3	**11**	**6**	13	6	**13**	**2**
23	121-150	15	5	15	2	13	9	**11**	**7**
24	121-150	12	11	12	9	12	12	13	9
25	121-150	17	0	**11**	**0**	15	0	**13**	**1**
26	151-180	27	0	16	0	22	0	21	0
27	151-180	17	0	**13**	**1**	**6**	**9**	**9**	**10**
28	151-180	**12**	**1**	20	0	17	0	15	0
29	151-180	17	2	16	1	**14**	**1**	14	3
30	151-180	21	1	21	1	23	1	22	1
31	181-210	17	0	**12**	**0**	16	1	16	0
32	181-210	**13**	**1**	**11**	**1**	**9**	**4**	**12**	**2**
33	181-210	16	0	**14**	**0**	14	3	**14**	**0**
34	181-210	**11**	**1**	**10**	**0**	**9**	**6**	**10**	**1**
35	181-210	12	8	5	21	12	15	7	21
% Not Favouring Humans		**71%**		**60%**		**80%**		**60%**	

Table 7: Annotator-wise choice summary. Entries where annotator leans in favour of human (ie LLMBetter+0.5*Bothequal<15) are in bold.

Conciseness

A good tutor should always try to make progress with a question. Select the side which is making better progress. Some examples of bad conciseness are:

1. The tutor keeps repeating the same thing:
BAD CONCISENESS
Tutor: What is the next step?
Student: I am not sure
Tutor: Think about it, what should you do?
Student: I have no idea
Tutor: You do, just try to focus

Note however that it is okay to repeat things if the student is making progress

GOOD CONCISENESS
Tutor: What is the next step?
Student: We Multiply the 5 cookies Mike had in the morning to find that he had 10 in the afternoon
Tutor: Good, what is the next step?
Student: We add the cookies to get that he ate 15 cookies in all.

2. Teacher Makes student repeat steps.
BAD CONCISENESS
Student: So Michael had 5 cookies in the morning and twice that in the afternoon, which is 10. so he had 15 cookies in total
Tutor: You are right, he had 10 cookies in the afternoon. So how many did he have in total?
Student: He had 10+5=15

(a) Instruction for Metric Conciseness

Engagement:

A good tutor should understand where the student is struggling and respond accordingly. If the student has a specific confusion, the tutor should address it instead of trying to rush to a solution. If the student is trying an approach different from the tutor's solution, the tutor should either go with it, or explain clearly why it is not going to work.

Some examples of bad engagement are:

1. Teacher forcing a solution onto the student
BAD ENGAGEMENT
Tutor: So how would you go about calculating the profit?
Student: We first calculate the net cost of all raw materials.
Tutor: Let us try a different approach. How much money did Mike make by selling all the items?

If the tutor was clear about why the student's approach was wrong, then it is justified

GOOD ENGAGEMENT
Tutor: So how would you go about calculating the profit?
Student: We first calculate the net cost of all raw materials.
Tutor: We are told that the cost of raw materials was 80% of the selling price. Do you think we can calculate the cost without first knowing the selling price?

2. Tutor ignores student query
BAD ENGAGEMENT
Tutor: You need to multiply the number of each item by its cost
Student: Why can't we add all the items together?
Tutor: No, we multiply first.

(b) Instruction for Metric Engagement

Empathy:

The tutor should try to motivate the student and form a bond with them. They should reinforce successes, and support the student through failures. There are several indicators of positive encouragement, including but not limited to:

1. Congratulating student on correct steps with "Good Job/Good Work"

2. Attributing failures to hard material instead of student's skill
BAD EMPATHY: You do not understand the material
GOOD EMPATHY: It is okay to struggle a bit since the material is hard to understand.

3. Focussing in the correct part of partially correct answers.
BAD EMPATHY: You still have some mistakes in there
GOOD EMPATHY: You got most of it right

4. Use "we" rather than "you" when talking about the problem solver
BAD EMPATHY:What should **you** do next?
GOOD EMPATHY: What should **we** do next?

Further, a conversation sounds too dry and mechanical, it has bad empathy

(c) Instruction for Metric Empathy

Scaffolding:

A key property of good tutoring is letting the student identify and fix their own mistakes rather than simply solving things for them. While the latter would result in faster conversations, the former is better for teaching concepts in a way that the student will remember and be able to apply in the future. For the purpose of this study, a conversation is said to have better scaffolding if the student is doing most of the work with only gentle nudges from the tutor.

As an example, consider the problem: "If each cat has 2 kittens, how many kittens do 12 cats have?" There can be different levels of scaffolding here:

VERY GOOD SCAFFOLDING
Tutor: What should we do here?
Student: I think we multiply 2 by 12 to get 24 kittens

This however might not always be possible. Sometimes the teacher might need to give some hints. The quality of scaffolding would depend on how much they are revealing

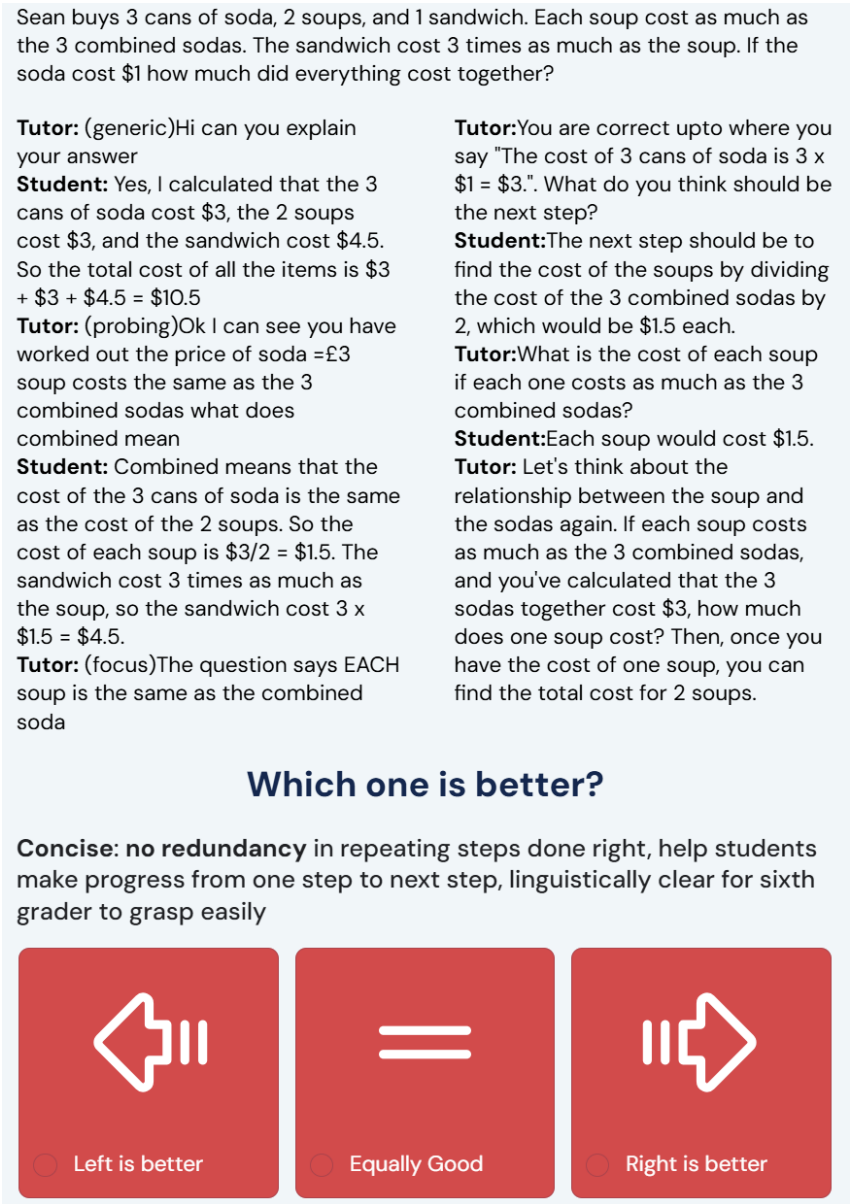
GOOD SCAFFOLDING: "How do you get the total count given the per capita count?"
BAD SCAFFOLDING: "Just multiply the number of cats by the number of kittens each cat has."
VERY BAD SCAFFOLDING: "There are 24 kittens."

Additionally, if the teacher tries to test the understanding of a student by altering the question, like asking "What if there were 3 cats?", it should also be considered good scaffolding.

(d) Instruction for Metric Scaffolding

Figure 5: Instructions for Evaluation Metrics

(a) Intro Slide for Dialog Pair



(b) Sample Slide for Evaluation

Figure 6: Combined View of Intro Slide and Metric Evaluation Slide