Can LLMs Effectively Simulate Human Learners? Teachers' Insights from Tutoring LLM Students

Daria Martynova[1] **Jakub Macina**[1,5] **Nico Daheim**[1,2]
Özge Nilay Yalçın[3] **Xiaoyu Zhang**[4,5] **Mrinmaya Sachan**[1]

[1]ETH Zurich [2]TU Darmstadt [3]Simon Fraser University
[4]City University of Hong Kong [5]ETH AI Center
{dmartynova, macinaj, ndaheim, mrinmaya}@ethz.ch
oyalcin@sfu.ca
xiaoyu.zhang@cityu.edu.hk

Abstract

Large Language Models (LLMs) offer many opportunities for scalably improving the teaching and learning process, for example, by simulating students for teacher training or lesson preparation. However, design requirements for building high-fidelity LLM-based simulations are poorly understood. This study aims to address this gap from the perspective of key stakeholders—teachers who have tutored LLM-simulated students. We use a mixed-method approach and conduct semi-structured interviews with these teachers, grounding our interview design and analysis in the Community of Inquiry and Scaffolding frameworks. Our findings indicate several challenges in LLM-simulated students, including authenticity, high language complexity, lack of emotions, unnatural attentiveness, and logical inconsistency. We end by categorizing four types of real-world student behaviors and provide guidelines for the design and development of LLM-based student simulations. These include introducing diverse personalities, modeling knowledge building, and promoting questions.

1 Introduction

Interactive student simulations provide a valuable tool for educators and students to prepare for lessons in a safe environment (Bradley and Kendall, 2014; McGarr, 2021; Chin et al., 2013) but often require substantial human resources, for example, for peer role-playing (Wang et al., 2021). Among other benefits, simulations allow pre-service teachers to practice guiding and managing students (Markel et al., 2023; McGarr, 2021), a skill they often feel unprepared for (Shank, 2023). In addition, in-service teachers can use simulations to enhance educational content and pedagogy (Aguilar and Kang, 2023). At the same time, students can benefit from learning by teaching a simulated peer (Chin et al., 2013). However, the need for human resources, e.g., to role-play students (Wang et al., 2021) or to set up mixed reality simulations (Aguilar and Telese, 2020), hinders a large-scale adaptation.

Simulating students using Large Language Models (LLMs) promises to alleviate this because LLMs can be accessed at any time and do not require involving vulnerable groups such as young learners. This is particularly attractive in educational settings, since frequent practice and exposure to diverse student behaviors are crucial to learning to teach effectively (Dagdag and Bandera, 2021; Loewenberg Ball and Forzani, 2009). Moreover, practicing with computer-simulated students reduces psychological strain from fear of making mistakes, among others (Chase et al., 2009). Finally, LLMs can offer personalized experiences by adapting to individual user needs and educational contexts (Eapen and Adhithyan, 2023) which has been shown to positively impact pre-service teacher training (Arnesen et al., 2019).

Specifically, we focus on the dialogue tutoring setting (Macina et al., 2023b), in which a human teacher is helping an LLM-simulated student to solve a problem. The goal of such a simulation is for the teacher to experience a realistic tutoring setting to improve their teaching skills.

To be useful, LLMs need to faithfully replicate real-world student behaviors, but the extent to which they can do so has not yet been explored well. In addition to more well-known shortcomings, such as their tendency to generate unnatural or false responses (Fu et al., 2024b; Tamkin et al., 2021), LLMs may be inconsistent with personal values (Kovač et al., 2024) and under-represent certain demographic groups when simulating personas (Wang et al., 2024a). Furthermore, a recent review highlighted that almost half of the studies that involved simulated learners did not validate whether their model was realistic enough to represent real students (Käser and Alexandron, 2024). This tendency raises questions about the reliability of these simulations in educational contexts. This

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 100–117

paper aims to address these concerns by answering the following research questions:

RQ1. How do LLM-simulated students deviate from the authentic behaviors of K12 students?

RQ2. How can LLM students be improved to better represent authentic student behaviors?

To answer these research questions, we conducted semi-structured interviews with 12 teachers who extensively interacted with LLM-simulated students during the creation of a dialogue tutoring dataset MathDial (Macina et al., 2023a). We used an analysis of this dataset to design interview questions based on two frameworks: the Community of Inquiry (CoI) (Garrison, 2016), which describes learning in online environments, and the Scaffolding theory (Reiser, 2004), which provides guidelines for effective teaching. See Fig. 1 for an overview of our interview design and analysis.

Our results indicate that LLMs can replicate some of the behaviors of an attentive student but still lack authenticity and diversity. Participants noted that the LLM students' responses were too technical and complex, lacked emotional expression, and sometimes were logically inconsistent or overly involved. We compared these findings with real-life student behaviors, which we classified into four categories in terms of scaffolding support needed as well as cognitive and social presence. Grounded in the Community of Inquiry and Scaffolding frameworks, these four categories offer a framework for designing educational LLM systems. We use these results to provide guidelines for developing more realistic LLM student simulations, including introducing diverse student personalities, modeling gradual knowledge building, and promoting question-asking.

2 Related Work

2.1 AI-Simulated Students in Tutoring

Simulations of learners have been used for various purposes, including teacher preparation, peer learning, and system evaluation (VanLehn et al., 1994). For example, (Matsuda et al., 2007) examined whether a machine learning model can replicate how students learn to solve linear equations. However, many early simulations required significant effort, despite modeling narrow settings (Matsuda et al., 2015).

LLMs have made these simulations considerably more accessible. Recent applications include simulating students to assess the quality of automatically generated questions (Lu and Wang, 2024), or using LLMs as teachable agents for learning debugging (Ma et al., 2024). However, whether the resulting model is realistic enough to represent a real student is not fully understood. A survey (Käser and Alexandron, 2024) found that only 3% of the studies that simulate learners do a post-factum validation of their model. Moreover, there is a growing trend of not validating LLM outputs or relying on LLMs validating themselves (Shankar et al., 2024). In contrast, we base our work on first-hand insights of teachers communicating with LLM students, which provides a deeper understanding of the realism of these models.

Namely, we interviewed teachers who took part in the collection of an existing open-source dataset MathDial (Macina et al., 2023a). We chose this dataset over other educational datasets such as NCTE (Demszky and Hill, 2023), Bridge (Wang et al., 2024b), or TalkMoves (Suresh et al., 2022), because, to the best of our knowledge, it is the only publicly available dataset of interactions between real teachers and LLM-simulated students. Additionally, the MathDial dataset is enriched by teacher annotations such as realism ratings.

2.2 Believability of LLM Simulations

According to (Park et al., 2023), believable agents provide an illusion of life and present a facade of realism in the way they appear to make decisions and act of their own volition. One common approach to evaluating believability is to compare LLM-generated and real-world (Hämäläinen et al., 2023). In our work, we use a similar approach by contrasting the experiences of teachers with LLM simulations and real interactions.

What constitutes a believable simulation is often dependent on its context; for example, applications in psychology focus on personal experience (Chen et al., 2023), while character motivation is important in games research (AlJammaz et al., 2024). In education, the focus is often on cognitive aspects, with the social component addressed in a too broad or unsystematic way (Jin et al., 2024; Jinxin et al., 2023). In this work, we also account for the social aspect by using the Community of Inquiry framework, which we introduce next.

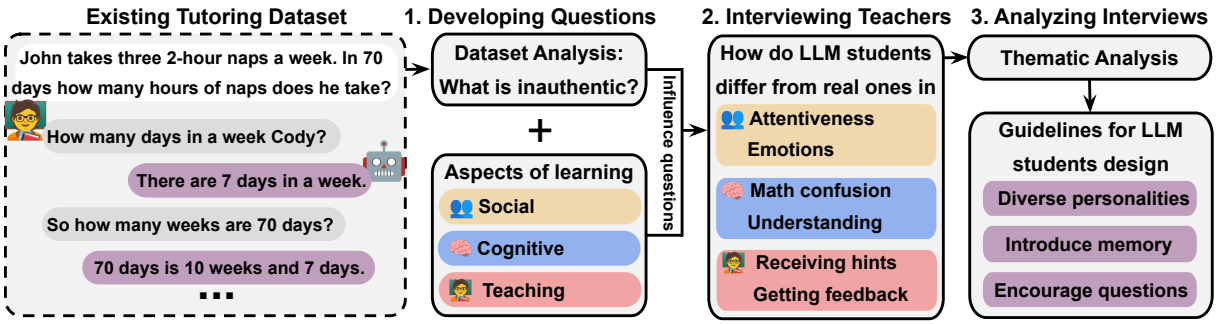

Figure 1: An illustration of our study stages: 1) We analyze an existing teacher-LLM tutoring dataset using the Community of Inquiry framework and derive interview questions from this analysis. 2) We interview teachers involved in data collection. 3) We outline guidelines for LLM student design and development.

2.3 Community of Inquiry and Scaffolding

Two important considerations in our study are the environment in which teachers use simulations and the form of teaching that is used. For the former, simulations are usually naturally used in an online setting, for example, through a web application. The Community of Inquiry (CoI) is a framework that is frequently used to understand online conversations in the context of education (Garrison, 2016). We adopt this framework to ground our interviews. CoI is based on three pillars: *social presence*, *cognitive presence*, and *teaching presence*. *Social presence* is defined as the ability of learners to project themselves socially and emotionally, thereby being perceived as "real people" in mediated communication (Garrison and Arbaugh, 2007). *Cognitive presence* is described in Garrison et al. (2001) as the extent to which learners are able to construct and confirm meaning through sustained reflection and discourse. *Teaching presence* is the design, facilitation, and direction of cognitive and social processes to achieve personally meaningful and educationally worthwhile learning outcomes (Garrison et al., 1999).

However, since the CoI framework gives limited attention to the active role of the teacher in guiding learning (Richardson and Lowenthal, 2017), we enriched the teaching presence with the Scaffolding theory (Wood et al., 1976; Quintana et al., 2004). In the setting of tutoring using scaffolding, the teacher guides the students and allows them to cognitively engage with the problem. Teachers usually follow a set of teaching strategies or moves (VanLehn, 2011; Nye et al., 2014; Hennessy et al., 2016) such as questioning with various effectiveness on learning (Michaels et al., 2008; Hennessy et al., 2016). The level of scaffolding needed depends on the student (Quintana et al., 2004; VanLehn, 2011) and often includes actively engaging them with the problem, including failure, which is more productive for learning (Kapur and Bielaczyc, 2012). In our paper, we investigate how the behavior of LLM-simulated students influences teaching strategies.

3 Methods

To answer RQ1, we focus on the existing open-source dataset MathDial (Macina et al., 2023a), in which teachers helped LLM-simulated students to solve a math problem, as shown in Fig. 1. To understand teachers' perceptions of LLM students' realism, we conducted interviews with participants of the MathDial study, described in Section 3.1. We then describe how analyzing the MathDial dataset provided initial insights into the realism of LLM student simulations (Section 3.2) and informed the development of interview questions (Section 3.3).

3.1 Participants

We recruited 12 teachers or tutors of STEM subjects among those who took part in the MathDial study (Macina et al., 2023a) through Prolific.[1] We pre-screened participants to ensure they taught technical subjects, aligning with experience in the study. After signing the consent form, each participant received as a reminder three example dialogues that they personally had in the MathDial study.

10 out of 12 participants teach mathematics, while the rest focus on natural sciences. The participants teach children and adolescents, in institutions ranging from primary schools to universities. 3 participants have been teaching for less than 3 years, while the rest — for more than 11 years. Most participants (8 out of 12) are UK-based, while the others work in Canada. 10 participants are female,

[1] https://www.prolific.com/

and the rest 2 are male, which is in line with the 80% proportion of female participants in the preceding study. The participants had an average of 35 dialogues with LLM students in the MathDial, with a standard deviation of 28. More details on participants' data can be found in Appendix A.

3.2 Developing Questions: MathDial Dataset Analysis

To design interview questions that capture teachers' perspectives on LLM students and address RQ1, we first analyzed the existing open-source tutoring MathDial dataset (Macina et al., 2023a), focusing on teachers' assessment of realism. In MathDial, teachers were asked to chat with a sixth-grade student simulated by an LLM and help them solve a math word problem. The LLM[2] was first prompted to generate an initial incorrect solution and then to act as a student who believes this solution is correct. The student persona was based on a name chosen from a culturally diverse set, a gender, and a specified type of confusion (see Macina et al. (2023a) for details). MathDial consists of 2861 dialogues produced by 90 participants, all of whom work as teachers. In addition to metadata such as teacher moves, each conversation is annotated by teachers with a rating on whether the interaction felt typical for a sixth-grade student, as well as optional open-ended "feedback about the conversation".

Topic Modeling of Teacher Feedback. We first analyze the open-ended feedback from teachers using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic modeling to find any concerns they had about LLM simulations. One of the recurring topics in the LDA analysis was "repetitive". By manual review, we found that 51 of the 377 feedbacks provided mentioned the student giving repetitive answers. Furthermore, 6 of 44 teachers who left feedback mentioned that it was frustrating for them when the student was stuck on the same solution.

Statistical Analysis of Teacher-assessed Realism. Here, we show a quantitative analysis of interaction realism ratings and the corresponding conversations. We focused on how conversations that the tutors rated as non-typical (21% of conversations) differed from those rated as typical (79% of conversations). We have performed statistical tests to check the independence of features when comparing typical and non-typical interactions. We chose features that are directly related to learning outcomes (e.g., the correctness of the final answer) or have the potential to impact student learning (e.g., emotions (Felten et al., 2006)). We used the Mann-Whitney U test (McKnight and Najab, 2010) for numerical features and the Chi-squared independence test (McHugh, 2013) for categorical features. Since we tested[3] multiple hypotheses, we used the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control for small p-values that occur by chance.

According to statistical tests, features whose distribution differed significantly among typical and non-typical interactions included the correctness of final student answer, conversation length, count of teacher moves revealing solution, and sentiment scores of teacher utterances computed using VADER (Hutto and Gilbert, 2014) (see Table 1). Conversations rated by teachers as non-typical were usually longer, less successful, and the solution was revealed more often. Sentiment scores of teacher messages were lower in non-typical interactions, while student sentiment remained similar for both types of conversations, leaning towards higher values. The unusual conversations might be more difficult for teachers as the students struggle to progress in their solutions. Detailed results of the statistical analysis can be found in Appendix C.

Comparative Analysis between Educational Datasets. A comparison between LLM-human dialogues and human-human conversational datasets showed that LLM students tend to be more active in conversation compared to real-life students. That is, we have compared MathDial and datasets with human-human interactions: 1) with transcripts from math classes (Suresh et al., 2022; Demszky and Hill, 2023) and 2) with text-based one-on-one tutoring dialogues in language learning (Caines et al., 2022; Stasaski et al., 2020). We computed dialogue metrics such as the total word count and the proportion of words contributed by teachers and students. The proportion of words in LLM-human dialogue is heavily skewed towards the LLM student, who contributes 68% of the total words. This contrasts sharply with human-to-human conversational data, where students typically account for only 12% to 34% of the word count.

[2] gpt-3.5-turbo, accessed through the OpenAI GPT-3 API [gpt-3.5-turbo]; available at: https://platform.openai.com/docs/models/gpt-3-5-turbo

[3] The analysis was done in Python using SciPy (Virtanen et al., 2020) and statsmodels (Seabold and Perktold, 2010) libraries. The significance level was set at 0.05.

Table 1: Comparison of conversations rated by teachers as typical or not.

Statistic of conversational dynamics	Typical interactions	Non-typical interactions
Proportion of dialogues	79%	21%
Success rate in resolving confusion	83%	43%
Average dialogue length (in turns)	12 ± 5.4	16 ± 6.4
Average frequency of teachers revealing solution	0.14 ± 0.18	0.22 ± 0.2
Average sentiment score of teacher messages	0.15 ± 0.3	0.1 ± 0.29
Average sentiment score of student messages	0.18 ± 0.32	0.17 ± 0.31

3.3 Interview Procedure and Questions

All the interviews were held online and lasted 1 hour. The interviews started with a warm-up task, in which participants were consecutively shown two short tutoring dialogues and were asked to distinguish whether the student responses were written by a human or an AI. This exercise served as an introduction to the interview topic: comparing interactions with real and simulated students. The main part of the interview focused on the experiences participants themselves had when communicating with LLM students in the MathDial study.

We developed the interview questions from the Community of Inquiry and Scaffolding frameworks (Section 2.3) and the MathDial data analysis (Section 3.2). We iteratively refined the interview questions based on team discussions and feedback from pilot interviews. We finalized a set of 9 questions prompting teachers to reflect on how their real students differ from LLM students. Questions related to social presence explored the attentiveness of students and their emotions, as LLM students tend to be repetitive and show higher sentiment scores. Questions from the cognitive presence category were motivated by observed deviations in LLM students' learning and focused on students' confusion, understanding, and solutions complexity. Finally, to address teaching presence, we asked about teachers' strategies, especially scaffolding and giving feedback, as teachers resorted to telling parts of the solution when the LLM student behaved unusually. The full list of interview questions and the rationale behind them can be found in Appendix B.

To summarize the discussion of each question, participants were asked to answer a 5-point Likert scale question assessing interactions with LLM students, e.g., realism of their emotions (see Fig. 2). After the interview, participants were reimbursed 34 USD per hour. The research was approved by the university Ethics Committee (EK-2024-N-6).

The interview data was analyzed using thematic analysis (Clarke and Braun, 2021). The initial coding was done by the main author, independently checked by two other team members, and iteratively refined. Finally, the codes were grouped into themes such as student emotions, language complexity, responsiveness, and demographics, as well as teachers' strategies and challenges.

4 Results

The main finding from the Likert scale survey answers (see Fig. 2) is that the LLM students did not authentically represent real human emotions. Apart from that, LLM students generally were able to simulate the learning process. Namely, aspects like teaching strategies, students' reactions to feedback, and math confusion were rated as more realistic. According to teacher ratings, LLM students were for the most part fairly attentive. In addition, the frequency of frustrating interactions and overly complicated solutions were rated relatively low.

Lack of Emotional Responses from LLM-simulated Students. 8 out of 12 participants noted that they did not seem to get particularly emotional responses from the LLM student. All teachers except one speculated that this perceived emotionlessness might just be the result of communication being only text-based and not being able to read the body language of the student.

To half of the participants, the student messages felt overall positive, with occasional emotions such as gratitude or relief. However, when asked about the common emotions of their real-life students when confused, all teachers primarily named negative ones such as frustration, fear, or embarrassment. A couple of participants believe their students react with denial, which LLMs did not portray: *'a human student is not going to immediately abandon a solution they've come up with.' (P11).*

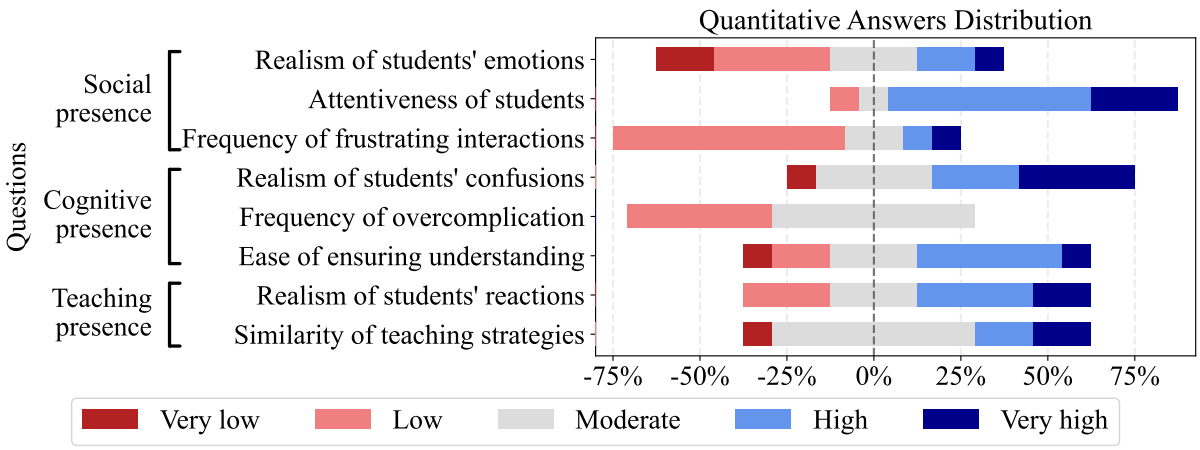

Figure 2: 5-point Likert scale ratings by teachers to questions about interactions with LLM-simulated students.

Eight participants mentioned that some of their students tend to give up or become quiet when they don't know how to solve a problem. LLM students failed to show this behavior: *'There weren't any students that just said, "Forget it. I can't do it. I give up." There was always a reattempt.' (P02).*

High Attentiveness. 10 out of 12 participants agreed that the LLM students felt rather attentive in the conversations. With real-life students, teachers see more diverse behaviors, e.g., *'You will have some children that are incredibly attentive, whereas, ... there are some children who have got very little interest in being there.' (P12).* Although LLM students generally resembled engaged students, P05 highlighted a difference: LLM students *'didn't ask any ... questions to help their understanding or make links with other things'.*

Inconsistent Behavior over Multiple Interactions. Two-thirds of participants pointed out that sometimes the LLM student felt like they were not following previous conversation. However, just as many teachers stressed that they are used to their students going off tangent, e.g., *'They always contradict themselves, and always say random things. And so that's not unusual at all.' (P01).*

Complex and Verbose Language Use by LLM-simulated Students. One of the frequently mentioned properties of LLM students which did not feel human-like to participants was the high language complexity. Also, two teachers noted how math formulas were extensively used by LLM students, which did not feel authentic. One participant highlighted how this hindered ensuring student understanding, as in real teaching students don't rely on *'mathematical language necessarily, they would actually talk to you in words.' (P09).*

Adaptation of Teaching Strategies for Interactions with LLM Students. Teachers have to adapt to the pace of their students; therefore, they pay high attention to the process of student learning, and they find several differences between LLM and actual students.

All participants emphasized the importance of scaffolding by breaking the problem down into smaller steps, as well as trying to give hints and not reveal parts of the solution. However, a quarter of participants noted that these approaches sometimes had to be adjusted when talking to LLM students, namely, teachers had to resort to telling parts of the solution. Two participants supposed that LLM-simulated students might have struggled because *'rather than try and take a step at a time, they were trying to solve everything altogether.' (P01).*

In MathDial, participants also frequently used approaches such as asking questions, finding other ways to solve a problem, and repeating. Participants found it to be *'no different to real life: you often have to repeat things and, if someone doesn't appear to understand how you said something the first time, you have to rephrase it.' (P05).* For P11, the experience of communicating with LLM felt *'analogous to working with humans: if your instructions are bad, your results are bad. ... as we ... learn more about how AI works, we are kind of also learning how humans work.' (P11).*

The Influence of Context on the Perception of Interactions with LLM-simulated Students. The participants teach students from different backgrounds, and some of their opinions on LLM students are also influenced by their diverse experiences. For example, P06 described that in some of the MathDial dialogues, *'That was interpretation*

where the gap was rather than actually a problem with the math. A common issue, actually, because a lot of our students ... have dyslexia'. Other teachers also mentioned dyscalculia, being non-verbal and having other special education needs, or having English not as their first language.

Differences in perception of interactions with LLM students could also be caused by the settings in which the participants teach. For example, P05 primarily works as a tutor and commented about LLM-simulated students: *'They seem to demonstrate a good growth mindset. That was probably quite different with students ... I work with, because it's one-to-one tuition and a lot are lacking that confidence already.' (P05).*

Half of the participants compared students' behavior across different subjects, e.g.: *'I have taught many subjects, and the only ones that really results sometimes in sobbing is math. ... Math can really trigger deep, deep emotions.' (P11).*

5 Discussion

5.1 Guidelines for LLM Students Design: Four Behavior Types

As our RQ1 aims to assess how believable the LLM student simulations are, we identify different groups of student behaviors in real life. Specifically, we do this based on the CoI framework and Scaffolding theory. In real-life education, some students need more scaffolding support, which means that the teacher provides step-by-step guidance to them and needs to engage them more actively in the process. Other students are more independent and actively participate in the problem-solving activity. Within both of these groups, we more specifically examine the social and cognitive presence of the students. That is, social presence relates to behaviors that help students engage and interact with the tutor, including demonstrating emotional expressiveness. On the other hand, cognitive presence focuses on how students process information, solve problems, and build knowledge. Table 2 provides an overview of behaviors not captured by LLM students for each category, as well as the importance participants placed on these issues and our proposed solutions, thereby addressing RQ2.

High Scaffolding Needs and Social Presence. Most of the interviewees agreed that LLM-simulated students were too engaged in conversations. We suggest that such simulations should have varying customizable levels of engagement, much as real students would. Sometimes, the simulated student might even stay silent or lose interest and attention, which could also give a valuable reason for teachers to self-reflect on the quality of teaching (Markel et al., 2023).

Participants often found the language used by LLM students to be too complex, lengthy, and technical, especially for children. Therefore, we propose having more variations in language complexity, intentionally regulating the length and formality of responses. Other suggestions include introducing grammar, spelling, or punctuation mistakes and, in the case of mathematics, limiting notations and the rigor of equations.

In addition to these behavioral tendencies, LLM students lacked emotional responses, especially the more negative ones: frustration, fear, or embarrassment. We propose to model a diverse range of student personalities, which in turn would lead to a diverse representation of emotions (Rusting and Larsen, 1997; Santos, 2016). A popular approach to portraying personalities is the Big Five theory (Costa and McCrae, 1999) which is also widely used in the development of LLMs (Jiang et al., 2024; Liu et al., 2024). This method of modeling diverse personalities might also broader represent previously mentioned engagement levels (Donovan et al., 2020; Zhang et al., 2020).

High Scaffolding Needs and Cognitive Presence. The way in which some LLM students' cognitive processes worked seemed unrealistic to our participants: their knowledge sometimes did not build gradually but made huge jumps. This is not only unrealistic, but it deprives teachers of practicing a recognized approach to teaching: leveraging the zone of proximal development (Vygotsky, 1978). The study (Jin et al., 2024) also focused on this limitation of LLMs and modulated the knowledge state as the conversation progressed, which could also be used in the setting of our research. One improvement we suggest future works to integrate is knowledge tracing (Scarlatos et al., 2025; Fu et al., 2024a) which is commonly used to estimate student knowledge and predict their responses. Another aspect that could be modeled to resemble human learning is forgetting information over time (Zhong et al., 2024).

Low Scaffolding Needs and Social Presence. Another behavior that LLM students failed to represent was asking questions. This meant that teachers had more control over the discussion flow, which is not always the case in real life. Jin et al. (2024) pro-

Table 2: Real-life student behavior LLMs failed to show and suggested solutions.

	High scaffolding needs	**Low scaffolding needs**
Social presence	■ Writing simple and short ■ Having negative emotions, being disengaged 🔧 *Introducing diverse personalities*	■ Asking questions 🔧*Promoting question-asking*
Cognitive presence	■ Gradual knowledge-building 🔧*Introducing memory*	● Disagreeing with teacher ● Changing tactic based on feedback 🔧 *No interventions needed*

■ Human-simulation gap ● Realistic behavior

poses a way to address this in the case of using simulated LLM students in the learning-by-teaching scenario. That is, their solution was to switch to the mode of asking questions with a period of three messages. We propose to use a similar technique that is more context-aware.

Low Scaffolding Needs and Cognitive Presence. Some students of our participants react with denial when told that their solution is wrong. In contrast, LLM students sometimes agree too readily with the teacher, completely changing their approach. This tendency of LLMs is called sycophancy bias (Perez et al., 2023) and originates from LLMs designed to follow instructions. Although this is useful in many contexts, when practicing interactions with a student, it is beneficial to put the effort into finding the correct method together.

Our participants sometimes observed that the LLM student was stuck on the same math problem solution, which was mostly recognized as common student behavior. This is in line with previous research, as LLMs are prone to being more stubborn when discussing mathematics than subjective topics (Ranaldi and Pucci, 2023). Dealing with students who struggle to progress is important for teachers; therefore, we do not recommend eliminating such types of interactions.

Practical Application Example. We propose that designers of LLM student simulations adopt a profile-oriented design approach (Jin et al., 2025; Wolff and Seffah, 2011), which involves incorporating diverse student personality traits and learning behaviors described in Table 2. Teachers could first pick a specific profile type of a simulated student, as well as their learning pace and knowledge level of a given topic. Using a base-prompt, a specific chatbot could be created for the teachers to interact with. A post-generation prompt could be used to make the final utterance shorter and simpler. This approach could increase the diversity of simulated student behaviors while ensuring consistency and realism, thereby making the simulations more inclusive and valuable for teacher practice.

5.2 Teacher Perceived Limitations of LLM Students

An overall trend we observed during the analysis was that LLMs mainly represented only certain student types and behaviors, depriving participants of richer teaching experiences. LLMs indeed have a tendency to portray an averaged representation of the data they were trained on. Our suggestion is to rather evaluate models by simulating the spectrum of student personas to allow for a more comprehensive teaching experience.

While LLMs often portrayed attentive students, some participants felt they resembled students with more surprising traits such as having learning challenges like dyslexia. We propose that LLM simulations should have the option to configure the simulated context, allowing teachers to get more valuable experience.

6 Conclusion

In this paper, we investigate the effectiveness of LLMs in simulating real K12 student behaviors by gathering insights from teachers who have tutored LLM-simulated students. Our findings reveal that LLMs fall short in replicating properties inherent in real-life students: emotions, especially negative, rather simple language, and the steady pace of learning. We address this issue by proposing a categorization of real-life student behaviors based on the level of needed scaffolding and relation to cognitive or social presence, and assess the LLM performance in representing each category. This categorization could serve as a guideline for evaluating novel LLM models for student simulations, for example by including more diverse student behavior types. Addressing these issues could enhance the

effectiveness and realism of future LLM student simulations in education, ultimately making educational resources more accessible, affordable, and personalized for a broader population.

Limitations

Our study has several limitations that future work could address. First, the dataset we analyzed generated the student simulations with an older GPT-3.5-turbo model. Future work could explore the differences in how other LLMs simulate students. Interestingly, for some tasks, more advanced models might perform worse: e.g., in Milička et al. (2024) GPT-4 (OpenAI, 2023), when prompted to simulate a one-year-old, gave more correct answers to logical questions than GPT-3.5-turbo. Moreover, studies comparing different LLMs find that some are more sensitive to the phrasing of math problems (Opedal et al., 2024) or less capable of reflecting emotional states (Ishikawa and Yoshino, 2025).

Secondly, the demographics of the study participants were limited: most of the participants were from the UK and the majority were female. While the high proportion of female teachers in our study reflects trends in the teaching profession (Government data about the UK's different ethnic groups, 2024), we acknowledge the potential impact of gender on the study results. For example, Sun et al. (2024) has shown that gender could influence the perceived anthropomorphism of a simulated persona. Further work could conduct larger-scale studies with more diverse demographics to analyze these dynamics further.

Finally, we limited the study scope to mathematics. However, as our participants also highlighted, real-life students' behavior differs depending on the subject. Similarly, LLMs might have varying attitudes towards different subjects, e.g., GPT models exhibit more anxiety when talking about mathematics (Abramski et al., 2023). Exploring other subjects and educational contexts could provide a more comprehensive understanding of the use of LLMs in student simulation.

Acknowledgements

We thank Peng Cui for discussions and feedback on early versions. Jakub Macina is supported by ETH AI Center doctoral fellowship.

References

Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. 2023. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3):124.

Jair J Aguilar and Seokmin Kang. 2023. Innovating with in-service mathematics teachers' professional development: The intersection among mixed-reality simulations, approximation-of-practice, and technology-acceptance. *International Electronic Journal of Mathematics Education*, 18(4):em0750.

Jair J Aguilar and James A Telese. 2020. Perceptions and opinions of the usability of simulations in a mathematics methods course for elementary pre-service teachers. *Journal of Education and Practice*, 11(12).

Rehaf AlJammaz, Noah Wardrip-Fruin, and Michael Mateas. 2024. Navigating faction systems: Insights and recommendations for more believable npcs in video games. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, pages 1–11.

Karen T Arnesen, Charles R Graham, Cecil R Short, and Douglas Archibald. 2019. Experiences with personalized learning in a blended teaching course for preservice teachers. *Journal of online learning research*, 5(3):275–310.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Elizabeth Gates Bradley and Brittany Kendall. 2014. A review of computer simulations in teacher education. *Journal of Educational Technology Systems*, 43(1):3–12.

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Swedish Language Technology Conference and NLP4CALL*, pages 23–35.

Catherine C Chase, Doris B Chin, Marily A Oppezzo, and Daniel L Schwartz. 2009. Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of science education and technology*, 18:334–352.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Doris B Chin, Ilsa M Dohmen, and Daniel L Schwartz. 2013. Young children can learn scientific reasoning with teachable agents. *IEEE Transactions on Learning Technologies*, 6(3):248–257.

Victoria Clarke and Virginia Braun. 2021. Thematic analysis: a practical guide. *SAGE Publications Ltd*.

PT Costa and RR McCrae. 1999. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(01):1999.

Januard Deñola Dagdag and Milky Mae D Bandera. 2021. Understanding the factors that influence students' behavior: Key towards an effective teaching. *Pedagogi: Jurnal Ilmu Pendidikan*, 21(2):144–148.

Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

Ryan Donovan, Aoife Johnson, Aine deRoiste, and Ruairi O'Reilly. 2020. Quantifying the links between personality sub-traits and the basic emotions. In *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part II 20*, pages 521–537. Springer.

Joel Eapen and VS Adhithyan. 2023. Personalization and customization of llm responses. *International Journal of Research Publication and Reviews*, 4(12):2617–2627.

Peter Felten, Leigh Z Gilchrist, and Alexa Darby. 2006. Emotion and learning: feeling our way toward a new theory of reflection in service-learning. *Michigan Journal of Community Service Learning*, 12(2):38–46.

Lingyue Fu, Hao Guan, Kounianhua Du, Jianghao Lin, Wei Xia, Weinan Zhang, Ruiming Tang, Yasheng Wang, and Yong Yu. 2024a. Sinkt: A structure-aware inductive knowledge tracing model with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 632–642.

Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024b. From text to self: Users' perception of aimc tools on interpersonal communication and self. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

D Randy Garrison. 2016. *E-learning in the 21st century: A community of inquiry framework for research and practice*. Routledge.

D Randy Garrison, Terry Anderson, and Walter Archer. 1999. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, 2(2-3):87–105.

D Randy Garrison, Terry Anderson, and Walter Archer. 2001. Critical thinking and computer conferencing: A model and tool to assess cognitive presence. *American Journal of Distance Education*.

D Randy Garrison and J Ben Arbaugh. 2007. Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and higher education*, 10(3):157–172.

Government data about the UK's different ethnic groups. 2024. School teacher workforce. Accessed: 2024-09-09.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, culture and social interaction*, 9:16–44.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8(1), pages 216–225.

Shin-nosuke Ishikawa and Atsushi Yoshino. 2025. Ai with emotions: Exploring emotional expressions in large language models. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 614–627.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach ai how to code: Using large language models as teachable agents for programming education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Shi Jinxin, Zhao Jiabao, Wang Yilei, Wu Xingjiao, Li Jiawen, and He Liang. 2023. Cgmi: Configurable general multi-agent interaction framework. *arXiv preprint arXiv:2308.12503*.

Manu Kapur and Katerine Bielaczyc. 2012. Designing for productive failure. *Journal of the Learning Sciences*, 21(1):45–83.

Tanja Käser and Giora Alexandron. 2024. Simulated learners in educational technology: A systematic literature review and a turing-like test. *International Journal of Artificial Intelligence in Education*, 34(2):545–585.

Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2024. Stick to your role! stability of personal values expressed in large language models. *PloS one*, 19(8):e0309114.

Zhengyuan Liu, Stella Yin, Geyu Lin, and Nancy Chen. 2024. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642.

Deborah Loewenberg Ball and Francesca M Forzani. 2009. The work of teaching and the challenge for teacher education. *Journal of teacher education*, 60(5):497–511.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 16–27.

Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the ai era? using llms as a teachable agent for debugging. In *International Conference on Artificial Intelligence in Education*, pages 265–279. Springer.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. 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.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Noboru Matsuda, William W Cohen, and Kenneth R Koedinger. 2015. Teaching the teacher: tutoring simstudent leads to more effective cognitive tutor authoring. *International Journal of Artificial Intelligence in Education*, 25:1–34.

Noboru Matsuda, William W Cohen, Jonathan Sewall, Gustavo Lacerda, and Kenneth R Koedinger. 2007. Predicting students' performance with simstudent: Learning cognitive skills from observation. *Frontiers in Artificial Intelligence and Applications*, 158:467.

Oliver McGarr. 2021. The use of virtual simulations in teacher education to develop pre-service teachers' behaviour and classroom management skills: implications for reflective practice. *Journal of Education for Teaching*, 47(2):274–286.

Mary L McHugh. 2013. The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

Sarah Michaels, Catherine O'Connor, and Lauren B Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27:283–297.

Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *Plos one*, 19(3):e0298522.

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24:427–469.

Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, Abulhair Saparov, and Mrinmaya Sachan. 2024. Do language models exhibit the same cognitive biases in problem solving as human learners? In *Proceedings of the 41st International Conference on Machine Learning*, pages 38762–38778.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In

Findings of the Association for Computational Linguistics: ACL 2023, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Chris Quintana, Brian Reiser, Elizabeth Davis, Joseph Krajcik, Eric Fretz, Ravit Duncan, Eleni Kyza, Daniel Edelson, and Elliot Soloway. 2004. A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences*, 13:337–386.

Leonardo Ranaldi and Giulia Pucci. 2023. When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.

Brian J. Reiser. 2004. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 13(3):273–304.

Jennifer C Richardson and Patrick Lowenthal. 2017. Instructor social presence: Learners' needs and a neglected component of the community of inquiry framework. In *Social Presence in Online Learning*, pages 86–98. Routledge.

Cheryl L Rusting and Randy J Larsen. 1997. Extraversion, neuroticism, and susceptibility to positive and negative affect: A test of two theoretical models. *Personality and individual differences*, 22(5):607–612.

Olga C Santos. 2016. Emotions and personality in adaptive e-learning systems: an affective computing perspective. *Emotions and personality in personalized services: Models, evaluation and applications*, pages 263–285.

Alexander Scarlatos, Ryan S. Baker, and Andrew Lan. 2025. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th Learning Analytics and Knowledge Conference, LAK 2025, Dublin, Ireland, March 3-7, 2025*. ACM.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: econometric and statistical modeling with python. *SciPy*, 7(1).

Melissa K Shank. 2023. Novice teachers' training and support needs in evidence-based classroom management. *Preventing School Failure: Alternative Education for Children and Youth*, 67(4):197–208.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

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.

Zhida Sun, Manuele Reani, Yunzhong Luo, and Zhuolan Bao. 2024. Anthropomorphism in chatbot systems between gender and individual differences. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *Preprint*, arXiv:2102.02503.

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

Kurt VanLehn, Stellan Ohlsson, and Rod Nason. 1994. Applications of simulated students: An exploration. *Journal of artificial intelligence in education*, 5:135–135.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Lev Semenovich Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024a. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024b. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199.

Xu Wang, Meredith Thompson, Kexin Yang, Dan Roy, Kenneth R Koedinger, Carolyn P Rose, and Justin Reich. 2021. Practice-based teacher questioning strategy training with elk: A role-playing simulation for eliciting learner knowledge. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27.

Dan Wolff and Ahmed Seffah. 2011. Ux modeler: a persona-based tool for capturing and modeling user

experience in service design. In *IFIP WG 13.2 Workshop at INTERACT 2011*, pages 7–16.

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.

Xiaojie Zhang, Guang Chen, and Bing Xu. 2020. The influence of group big-five personality composition on student engagement in online discussion. *International Journal of Information and Education Technology*, 10(10):744–750.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(17), pages 19724–19731.

A Participant Information

Table 3: Participant demographics, teaching experience, and the number of dialogues with LLM students in MathDial (Macina et al., 2023a).

ID	Age	Gender	Country	Student Ages	Subjects	Teaching Experience	#Dialogues
P01	40–49	Female	UK	5–9	Primary school subjects, including mathematics	15+ years	50
P02	40–49	Female	Canada	10–14	Mathematics	11–15 years	40
P03	30–39	Female	UK	0–9	Primary school subjects, including mathematics	1–3 years	100
P04	40–49	Female	UK	5–9	Primary school subjects, including mathematics	15+ years	70
P05	30–39	Female	UK	5–17	Mathematics, computer science, literature	11–15 years	19
P06	40–49	Female	UK	18+	Environmental science	15+ years	35
P07	20–29	Female	Canada	5–14, 18+	Mathematics, chemistry	1–3 years	30
P08	40–49	Male	UK	18+	Applied statistics	15+ years	20
P09	50–59	Female	UK	10–17	Mathematics, English as a foreign language, literature	15+ years	25
P10	20–29	Female	Canada	5–17	Biochemistry, English as a foreign language	1–3 years	10
P11	50–59	Female	Canada	5–17	Mathematics, computer science	15+ years	10
P12	40–49	Male	UK	5–14	Primary school subjects, including mathematics	11–15 years	5

B Interview Questions

Table 4: Interview questions and their connection to preceding MathDial analysis and theoretical frameworks: Community of Inquiry (CoI) (Garrison, 2016) and Scaffolding (Reiser, 2004)

	Qualitative and Quantitative Questions	Rationale
1	**Question:** In MathDial, how *attentive* were the students? **Probes:** Did it seem like the student was following what you were saying? If not, what were the examples when the student seemed like they didn't follow you? Were there cases when the student contradicted themselves? How do these cases compare to your real life experience? **Evaluation:** How attentive the MathDial students felt like? 1 (Not at all) - 5 (Extremely)	**MathDial analysis:** Some participants mentioned in the feedback field that the student's messages were repetitive **CoI framework:** Social presence
2	**Question:** How *engaged* are your students in math problem discussions? **Probes:** How much do they participate in conversation? How does it compare with the dialogues you had in the study? **Evaluation:** How engaged were the MathDial students? 1 (Much less than your students) - 5 (Much more than your students)	**MathDial analysis:** Compared to human-human educational datasets, the student in MathDial talks much more **CoI framework:** Social presence
3	**Question:** Which interactions with MathDial students were *frustrating* for you? **Probes:** How similar were they to the real life teaching? How do you deal with these? **Evaluation:** How often were MathDial interactions frustrating? 1 (Never) - 5 (Almost always)	**MathDial analysis:** The participants answers tend to have lower sentiment scores in conversations where the student interactions are perceived as non-typical **CoI framework:** Social presence
4	**Question:** Did you adjust your *teaching strategies* in MathDial? **Probes:** For example, how did you balance giving hints and giving parts of the solution? How do you do it in your real life teaching? **Evaluation:** How similar to real life were your teaching strategies in MathDial? 1 (Not at all) - 5 (Extremely)	**MathDial analysis:** The teachers tended to more often reveal part of the solution in conversations with non-typical interactions **Theoretical framework:** Scaffolding theory and Teaching presence from CoI
5	**Question:** What *feedback* do you give your students? **Probes:** How do they typically react to it? Were the student's reactions to feedback in MathDial similar to the typical reaction of your students? **Evaluation:** How realistic were students' reactions to feedback in MathDial? 1 (Not at all) - 5 (Extremely)	**MathDial analysis:** There was a cap on the number of messages teachers could send, so the feedback might have been rather limited **CoI framework:** Teaching presence

Table 4: Interview questions and their connection to preceding MathDial analysis and theoretical frameworks: Community of Inquiry (CoI) (Garrison, 2016) and Scaffolding (Reiser, 2004)

	Qualitative and Quantitative Questions	Rationale
6	**Question:** What *emotions* are common to your students due to math confusion? **Probes:** How closely was it represented in the MathDial study? How do you behave when the students convey emotions you listed? **Evaluation:** How realistic were students' emotions in MathDial? 1 (Not at all) - 5 (Extremely)	**MathDial analysis:** Sentiment score of student utterances is distributed independently of how typical the student interactions were **CoI framework:** Social presence
7	**Question:** What was the common *reason of confusion* in MathDial? **Probes:** How does it align with most common issues your students have? **Evaluation:** How realistic was students' confusion in MathDial? 1 (Not at all) - 5 (Extremely)	**MathDial analysis:** Some teachers assessed student's confusion as non-typical **CoI framework:** Cognitive presence
8	**Question:** In real life teaching, how do you ensure the *concept understanding*? **Probes:** What do you usually do after the correct solution was found? Do you continue the problem discussion? If yes, how? **Evaluation:** It was easy to ensure understanding of students in MathDial 1 (Strongly disagree) - 5 (Strongly agree)	**MathDial analysis:** Mainly the teachers stopped the dialogue after the student has found the correct solution **CoI framework:** Cognitive presence
9	**Question:** In real life teaching, how do you handle *overcomplicated solutions*? **Probes:** For example, do you let them explore their solution further? Or do you try to guide them to an easier solution? **Evaluation:** How often were MathDial solutions overcomplicated? 1 (Never) - 5 (Almost always)	**MathDial analysis:** LLM students sometimes used more complex methods (e.g., introducing variables) when the problem could be solved without them **CoI framework:** Cognitive presence

C Statistical Tests on MathDial

Table 5: Results of statistical tests comparing distribution of numerical features in typical and non-typical interactions in MathDial. U-statistic (McKnight and Najab, 2010) and p-value adjusted using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) are provided, with significant results (adjusted p-value < 0.05) marked with an asterisk (*).

(a) Teacher-annotated and sentiment features

Feature	U-statistic	Adjusted p-value
Teacher-assessed cognition of LLM student		
Confusion authenticity	220357	7.47e-145*
Step of first error in solution	74669	7.02e-01
Counts of teacher-annotated teacher moves		
Revealing parts of solution	876991	6.93e-36*
Constraining to make progress	790520	3.75e-12*
Talking casually	600816	7.49e-04*
Generalizing aspects of problem	721417	3.52e-03*
Teacher sentiment scores		
Mean	605884	3.52e-03*
Median	605894	3.52e-03*
Minimum	606569	3.52e-03*
Standard deviation	620603	3.62e-02*
Maximum	631284	1.46e-01
LLM student sentiment scores		
Minimum	615997	1.77e-02*
Maximum	690558	2.97e-01
Mean	653972	7.41e-01
Median	655628	7.98e-01
Standard deviation	661922	8.96e-01

(b) Interaction and problem-related metrics

Feature	U-statistic	Adjusted p-value
Conversation characteristics		
Number of turns	920056	5.24e-46*
Conversation index	685230	4.61e-01
Ground-truth solution characteristics		
Number of words	638996	3.04e-01
Number of steps	650522	6.35e-01
Math problem characteristics		
Order of the problem in session	648169	6.81e-01
Identifier	652030	7.02e-01
Sentiment score	660511	8.98e-01
Number of words	669497	8.98e-01
Arithmetic operation percentages in solution		
Addition	701925	7.25e-02
Subtraction	676748	6.73e-01
Multiplication	652588	6.73e-01
Division	663954	9.77e-01

Table 6: Results of statistical tests comparing distribution of categorical features in typical and non-typical interactions in MathDial. χ^2 statistic (McHugh, 2013) and p-value adjusted using Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) are provided, with significant results (adjusted p-value < 0.05) marked with an asterisk (*).

Feature	χ^2 statistic	Adjusted p-value
Teacher-assessed cognition of LLM student		
Correctness of final answer	479.83	1.28e-103*
Error category (calculation or conceptual)	6.38	6.35e-01
Teacher and LLM student data		
Teacher identifier	358.66	3.74e-33*
Student's name (from prompt)	40.82	3.55e-02*
Student's math struggle type (from prompt)	9.56	1.97e-01
Student's gender (from prompt)	0.81	6.35e-01
Topics mentioned in math problem		
Time	0.15	8.68e-01
Percent	0.09	8.96e-01
Money	0.07	8.96e-01
Age	0.03	8.96e-01
Fractions	0.04	8.96e-01