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

Table of Contents

1 Participant Information

Table 1: Participant demographics, teaching experience, and the number of dialogues with LLM students in MathDial [3].

ID	Age	Gender	Country of Residence	Student Ages	Subjects	Teaching Experience	MathDial Dialogues Count
P01	40–49	Female	UK	5–9	Primary school subjects, including mathematics	More than 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	More than 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	More than 15 years	35
P07	20–29	Female	Canada	5–14, 18+	Mathematics, chemistry	1–3 years	30
P08	40–49	Male	UK	18+	Mathematics (applied statistics)	More than 15 years	20
P09	50–59	Female	UK	10–17	Mathematics, English as a foreign language, literature	More than 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	More than 15 years	10
P12	40–49	Male	UK	5–14	Primary school subjects, including mathematics	11–15 years	5

2 Interview Questions

Table 2: Interview questions and their connection to preceding MathDial analysis and theoretical frameworks: Community of Inquiry (CoI) [2] and Scaffolding [6]

	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

Table 2: Interview questions and their connection to preceding MathDial analysis and theoretical frameworks: Community of Inquiry (CoI) [2] and Scaffolding [6]

	Qualitative and Quantitative Questions	Rationale
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
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

Table 2: Interview questions and their connection to preceding MathDial analysis and theoretical frameworks: Community of Inquiry (CoI) [2] and Scaffolding [6]

	Qualitative and Quantitative Questions	Rationale
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

3 Statistical Tests on MathDial

Table 3: Results of statistical tests comparing distribution of numerical features in typical and non-typical interactions in MathDial. U-statistic [5] and p-value adjusted using Benjamini-Hochberg procedure [1] 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 4: Results of statistical tests comparing distribution of categorical features in typical and non-typical interactions in MathDial. χ^2 statistic [4] and p-value adjusted using Benjamini-Hochberg procedure [1] 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

Bibliography

[1] Benjamini, Y., Hochberg, Y.: 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 (1995)

[2] Garrison, D.R.: E-learning in the 21st century: A community of inquiry framework for research and practice. Routledge (2016)

[3] Macina, J., Daheim, N., Chowdhury, S.P., Sinha, T., Kapur, M., Gurevych, I., Sachan, M.: Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. arXiv preprint arXiv:2305.14536 (2023)

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

[5] McKnight, P.E., Najab, J.: Mann-whitney u test. The Corsini encyclopedia of psychology pp. 1–1 (2010)

[6] Reiser, B.J.: Scaffolding complex learning: The mechanisms of structuring and problematizing student work. Journal of the Learning Sciences **13**(3), 273–304 (2004). `https://doi.org/10.1207/s15327809jls1303_2`