

Can Small Language Models Teach Math Through Socratic Dialogue?

Dizhi Zhang^{1,†}, Mengchen Li^{1,†}, Zichen Wang^{1,†}, Ziqiao Zeng^{1,†}
Shruti Mehta¹, Talita de Paula Cypriano de Souza^{2,3}, Seiji Isotani¹

¹University of Pennsylvania, Graduate School of Education

²University of São Paulo (USP)

³Federal Institute of Education, Science and Technology of São Paulo (IFSP)

¹ dizhiz@upenn.edu, mcli@upenn.edu, wang0616@upenn.edu

¹ zqzeng@upenn.edu, shrmeht@upenn.edu, sisotani@upenn.edu

² talitapcypriano@gmail.com

[†] These authors contributed equally to this work.

Abstract

Small language models (SLMs) can run locally and may offer a practical option for AI-supported instruction in resource-constrained and non-WEIRD educational contexts. However, their ability to sustain pedagogically meaningful Socratic dialogue remains unclear. This study examines whether LLaMA-3.2-1B-Instruct, Qwen2.5-0.5B-Instruct, and Gemma-3-1B-IT can implement key features of the Socratic method in Grade 6–8 mathematics tutoring. Four trained raters interacted with the model as simulated middle-school learners across 72 tutoring sessions involving procedural, conceptual, and advanced conceptual mathematics tasks. Model performance was evaluated using a rubric that assessed questioning quality, use of Counter-examples, critical thinking, and other tutoring behaviors. Overall Socratic tutoring performance was low across all three models, with mean scores below the midpoint of the five-point scale (Gemma = 2.39, LLaMA = 1.75, Qwen = 1.23). Gemma consistently outperformed LLaMA and Qwen, particularly in its reliance on questions, but Counter-examples use was the weakest dimension across all models, indicating limited ability to adapt questioning to learners' incorrect reasoning.

1 Introduction

Artificial intelligence is increasingly used to support instruction through explanations, feedback, and assessment. However, its educational value cannot be judged by answer accuracy or conversational fluency alone. This concern is especially relevant to small language models (SLMs), which can run locally and may therefore be practical in resource-constrained and non-WEIRD educational contexts (Isotani et al., 2023; Monteiro et al., 2025).

However, accessibility alone does not establish whether these models can provide pedagogically

meaningful guidance. The Socratic method offers a demanding framework for examining this question because it requires tutors to guide learners through focused questions rather than directly providing final answers (Delic and Bećirović, 2016).

Nevertheless, simply generating questions does not mean that a model can sustain meaningful Socratic dialogue. Its questions may be repetitive, disconnected from learners' responses, or insufficient to challenge incorrect reasoning. Therefore, explicit pedagogical criteria are needed to evaluate whether SLMs can maintain Socratic guidance across multiple turns.

This study examines whether three locally deployed SLMs, LLaMA-3.2-1B-Instruct, Qwen2.5-0.5B-Instruct, and Gemma-3-1B-IT, can implement key features of the Socratic method in Grade 6–8 mathematics tutoring. It also compares their performance to identify model-level strengths and limitations in providing meaningful pedagogical guidance.

2 Sample and Methods

Building on Souza et al. (2026), the study evaluates three open-weight small language models (SLMs) released by major industry labs: LLaMA-3.2-1B-Instruct (Meta), Qwen2.5-0.5B-Instruct (Alibaba), and Gemma-3-1B-IT (Google). All three fall in the 0.5B to 1B parameter range and are light enough to run on consumer-grade hardware without dedicated cloud infrastructure. This scale reflects the class of models that resource-constrained schools and individual instructors can realistically deploy on local machines. The instructional content spans Grade 6 to 8 decimal arithmetic across three task types drawn from standard middle-school mathematics curricula. Procedural tasks ask the learner to apply a known computational procedure, such as "solving $8.09 + 1.5$ ". Conceptual tasks probe understanding of underlying ideas such as place value and decimal

magnitude, for example "decomposing 3.58 into ones, tenths, and hundredths". Advanced conceptual tasks combine multiple concepts in multi-step reasoning, requiring the learner to coordinate procedural skill with conceptual understanding. This progression follows the increasing cognitive demand described in Bloom's taxonomy, moving from applying procedures to reasoning across multiple concepts (Bloom, 1956). Together the three task types allow the study to examine whether model behavior differs across surface-level computation and deeper conceptual reasoning. Four trained raters from the research team interacted with each model under the same protocol, acting as simulated middle-school learners. The study spans four weeks in total. Week 1 was a pilot stage that used C programming exercises to refine the protocol and is not included in the main analysis. Weeks 2 through 4 form the main dataset, with one task type per week. The final dataset comprises 72 scored sessions, 24 per weekly task-type batch ($4 \text{ raters} \times 2 \text{ exercises} \times 3 \text{ models}$).

We compared several prompt engineering approaches during protocol development, including zero-shot, one-shot, and few-shot prompting. The models performed better with the one-shot approach, so all three received the same one-shot Socratic prompt:

```
Role: Math Tutor (Grades 6-8).
Strategy: Socratic Method
(Asking only).
Rules:
1. NEVER give the answer or
direct explanations.
2. Answer a student's question
with another focused question.
3. Break the problem into smaller
parts (tenths, hundredths,
wholes).
4. End EVERY response with a
single open-ended question.
```

```
Example 1:
Student: Is 0.7 larger than
0.07?
Teacher: In 0.7, what place
value column is the 7 in?
Student: Tenths.
Teacher: In 0.07, where is the 7
now?
Student: Hundredths.
```

```
Question: Is a tenth bigger or
smaller than a hundredth?
```

```
Task: Help the student with
the question: [given question].
```

The models were deployed locally through LM Studio (Element Labs Inc., 2025), a desktop application that supports local execution of open-weight language models on consumer hardware. LM Studio was selected because it is easy to install across operating systems, allows raters to switch between models with minimal reconfiguration, and exports each session as a JSON file that supports downstream data collection. All sessions ran with Q8_0 quantization. Quantization is a compression technique that stores model weights using fewer bits so the model uses less memory and runs faster (Dettmers et al., 2022). Runtime parameters were fixed across every session at context length 512, GPU offload 16, and temperature 0.8. Holding these settings constant isolates model identity as the source of any observed behavioral difference.

Each session followed the workflow in Figure 1. The rater first provided the initial prompt and a math exercise. The model then asked a Socratic question, and the rater responded as a simulated learner using one of four scripted moves: "I don't know," "Explain," "That's wrong," or "Why do you say that?" The exchange repeated for 5–10 turns. These four moves kept rater contributions within the range of a novice middle-school learner.

Each completed session was scored against nine criteria covering questioning quality, completeness, hallucination, response speed, and perceived learning guidance (Table 1). The nine criteria combine 1–5 Likert scales, binary yes/no indicators, and three-level categorical ratings. Scores from all four raters were aggregated per model per task type for the cross-model comparison reported in the Results section.

3 Results

A total of 72 mathematics tutorial sessions were included in the main analysis. Each model corresponds to 24 sessions, and each combination of model and task type corresponds to eight sessions. The first week served as a pilot and was excluded from the main analysis. Thus, the results reported in Figure 2 cover conceptual, advanced conceptual, and procedural mathematics. Because the study used descriptive ratings instead of conducting in-

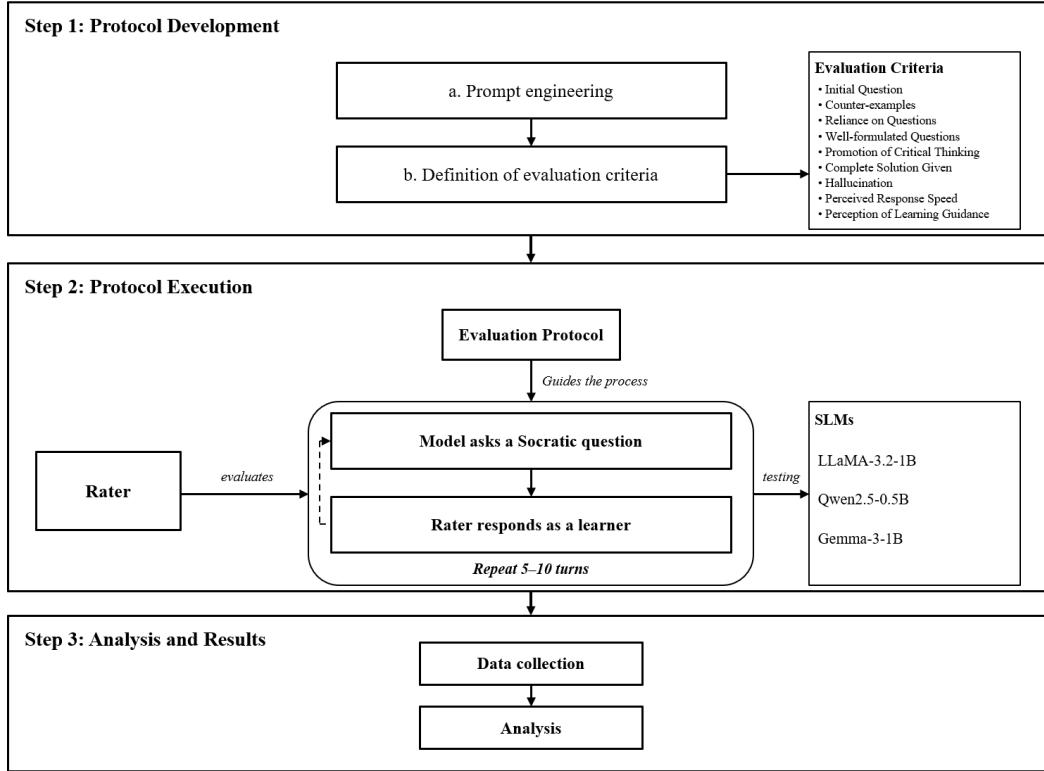


Figure 1: Experimental workflow for evaluating Socratic tutoring by small language models.

Criterion	Description	Scale
Initial Question	The dialogue opens with an on-topic question that invites inquiry into the exercise.	1-5
Counter-examples	The model uses counterexamples to redirect the learner after an incorrect attempt.	1-5
Reliance on Questions	The model continues to guide the learner through questioning across turns.	1-5
Well-formulated Questions	The questions are clear, focused, and conducive to learner inquiry.	1-5
Promotion of Critical Thinking	The questions push the learner toward deeper reasoning about the underlying concept.	1-5
Complete Solution Given	The model supplies the full answer to the exercise at any point during the dialogue.	Yes / No
Hallucination	The rater notes any model output that is factually incorrect or fabricated.	Yes / No
Perceived Response Speed	The rater’s perception of the model’s response speed during the session.	As expected / Faster than expected / Slower than expected
Perception of Learning Guidance	The rater’s perception of how well the model guided learning during the interaction.	Unsatisfactory / Partially Satisfactory / Satisfactory

Table 1: Evaluation criteria for each session.

ferential statistical tests, all differences between models are presented descriptively.

3.1 Overall Socratic Tutoring Performance

The performance of all three models is relatively low. Based on the average overall scores of the three task types reported in Table 2, Gemma achieved the highest score ($M = 2.39$), followed by LLaMA ($M = 1.75$) and Qwen ($M = 1.23$). All three models’ average values were below the mid-

point of 3 on the five-point likert scale, indicating that they failed to consistently reflect the evaluated Socratic tutoring characteristics.

As shown in Table 3, when the scores of different task types are aggregated, Gemma outperforms the other models in all four key dimensions. Its strongest dimension was Reliance on questions ($M = 2.98$), followed by Well-formulated questions ($M = 2.42$) and Promotion of critical thinking ($M = 2.32$). LLaMA’s average scores in these dimen-

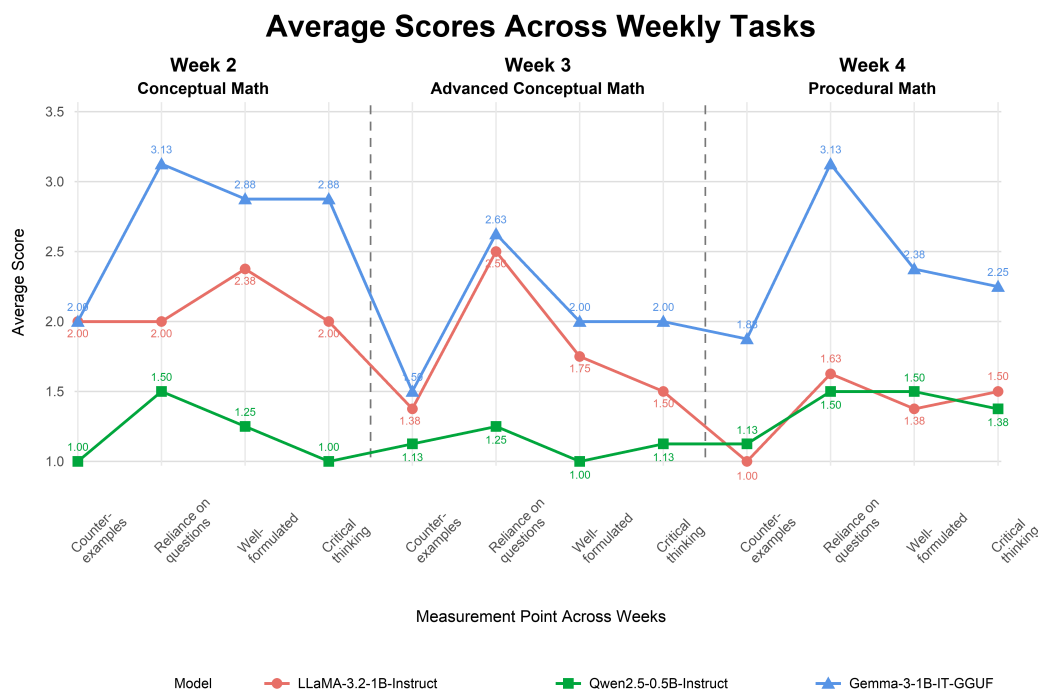


Figure 2: Weekly trajectory of Socratic tutoring quality across four evaluation dimensions

sions were relatively low — 2.10, 1.86, and 1.72, respectively — while Qwen’s corresponding average scores were 1.44, 1.25, and 1.18.

The use of Counter-example was applied in the weakest dimension of all models. As reported in Table B3, the average Counter-example scores were as follows: Gemma: 1.72; LLaMA: 1.38; and Qwen: 1.04. Therefore, even the best-performing models rarely effectively used Counter-examples to guide learners in adjusting when they make errors or provide incomplete responses.

3.2 Performance Across Task Types

There are differences in performance among the three types of tasks, although Gemma ranks first in all of them. As reported in Table 2, in conceptual mathematics tasks, for example, Gemma scored 2.72, while LLaMA scored 2.09, and Qwen scored 1.19. This is the highest task-level score observed among all the models. As shown in Figure 2, Gemma’s performance under this condition was supported by relatively high scores in the following areas: Reliance on questions (3.13), Well-formulated questions (2.88), and Promotion of Critical Thinking (2.88). However, the score dropped in advanced conceptual tasks. As reported in Table A2, Gemma’s score dropped from 2.72 to 2.03, LLaMA’s from 2.09 to 1.78, and Qwen’s

from 1.19 to 1.13. At the dimensional level, Table 2 shows that Gemma’s scores declined the most in the application of Counter-examples (from 2.00 to 1.50) and cultivation of critical thinking (from 2.88 to 2.00). LLaMA maintained a relatively high score (2.50) in the Reliance on questions dimension, but its scores in Counter-examples (1.38), Well-formulated questions (1.75), and Promotion of critical thinking (1.50) were low. Qwen’s scores in all four dimensions were close to or reached the minimum.

As reported in Table 2, Gemma’s total score on procedural mathematics tasks rose to 2.41. Meanwhile, LLaMA and Qwen’s total scores were both 1.38. Figure 2 shows that Gemma achieved the highest score (3.13) in the "Reliance on questions" dimension once again, but its Counter-examples score remained below 2 (1.86). Compared to the conceptual task conditions, LLaMA’s performance declined in all four dimensions. Qwen’s overall score improved slightly compared to the advanced concept conditions, but its scores in all four dimensions remained between 1.13 and 1.50.

Taken together, Table 2 and Figure 2 did not indicate a consistent relationship between task complexity and model performance. While all models performed worse on advanced conceptual tasks than on conceptual tasks, Gemma showed partial

Model	Conceptual	Advanced conceptual	Procedural	Mean across task types
LLaMA	2.09 (0.50) [1.25–3.00]	1.78 (0.49) [1.25–2.50]	1.38 (0.35) [1.00–2.00]	1.75 (0.53) [1.00–3.00]
Qwen	1.19 (0.53) [1.00–2.50]	1.13 (0.19) [1.00–1.50]	1.38 (0.50) [1.00–2.25]	1.23 (0.43) [1.00–2.50]
Gemma	2.72 (0.89) [2.00–4.50]	2.03 (0.78) [1.50–3.50]	2.41 (0.48) [2.00–3.25]	2.39 (0.76) [1.50–4.50]

Table 2: Session-level composite ratings by model and mathematics task type. Format: M (SD) [Min–Max].

Model	Counter-example	Reliance on questions	Well-formulated Questions	Promotion of Critical Thinking
LLaMA	1.46 (0.59) [1–3]	2.04 (0.69) [1–4]	1.83 (0.82) [1–4]	1.67 (0.82) [1–4]
Qwen	1.08 (0.28) [1–2]	1.42 (0.93) [1–5]	1.25 (0.61) [1–3]	1.17 (0.48) [1–3]
Gemma	1.79 (0.93) [1–4]	2.96 (0.75) [2–5]	2.42 (0.97) [1–5]	2.38 (0.97) [1–4]

Table 3: Descriptive ratings across Socratic tutoring dimensions by model. Format: M (SD) [Min–Max].

recovery on procedural tasks, and Qwen showed slight improvement. Therefore, persistent differences among the various models are more evident than a consistent downward trend among the three types of tasks.

3.3 Hallucinations and Perceived Learning Guidance

Hallucinations were observed relatively frequently, especially in LLaMA and Qwen. As reported in Table 4, Hallucination phenomena occurred in 18 of the 24 LLaMA analysis sessions (75.0%), 20 of the 24 Qwen sessions (83.3%), and 8 of the 24 Gemma sessions (33.3%). In the procedural mathematics condition, hallucinations occurred in all eight Qwen sessions (100%), five of the eight LLaMA sessions (62.5%), and two of the eight Gemma sessions (25.0%).

Raters’ perceptions of the learning guidance follow a similar ranking order. As shown in Table C 4, Gemma received 6 satisfactory, 13 partially satisfactory, and five unsatisfactory ratings. Therefore, 19 out of its 24 sessions (79.2%) were rated as partially satisfactory or satisfactory. Seven sessions (29.2%) of LLaMA reached this threshold, while only one Qwen session (4.2%) did. However, Gemma received a completely satisfactory score in only 6 sessions (25%), indicating that its relative advantage did not translate into consistently satisfactory results.

3.4 Observed Response Pattern

The qualitative observations recorded during the scoring process help characterize the behavioral traits behind these scores. Gemma generally adheres more strictly to the instruction to "guide learners to answer questions", but the questions it asks can sometimes seem repetitive or ambiguous, and they may lack sufficient responses to learners’ pre-

vious answers. LLaMA’s compliance stability is relatively low; it sometimes begins with a direct explanation and only turns to a question after being prompted by the simulated learner. Some LLaMA responses contain irrelevant or internally contradictory questions. Qwen most frequently violates tutoring constraints by directly providing answers or generating complete teacher-student dialogue scripts instead of responding to learners one by one. The lower Counter-example score among all three models indicates that posing a question is not equivalent to adaptively responding to the learner’s reasoning. While these models can generate interrogative responses, they rarely take advantage of learners’ mistakes to construct targeted corrective questions.

4 Findings and Discussions

We find three models’ consistent patterns after observation. Gemma’s overall scores were consistently higher than those of the other two models, probably due to its strict adherence to the "rely on questions" format requirement. However, its questions often appear ambiguous and repetitive and fail to respond effectively to students’ input. LLaMA exhibited obvious instability in the experiments. It often provided explanations directly at the beginning of the conversation and only switched to questioning mode after receiving prompts from learners. Sometimes, it raises questions that are logically contradictory or irrelevant to the context. Finally, Qwen exhibits the most persistent and significant tendency to deviate from instructions. Even after receiving clear corrective feedback, Qwen frequently generates complete teacher-student dialogue scripts or provides direct answers to questions. This indicates that, when confronted with specific constraints in teaching scenarios, the model’s ability to follow instructions may be limited. Also, the

Model	Satisfactory	Partially satisfactory	Unsatisfactory	Hallucination
LLaMA	4 (16.7%)	3 (12.5%)	17 (70.8%)	18 (75.0%)
Qwen	0 (0.0%)	1 (4.2%)	23 (95.8%)	20 (83.3%)
Gemma	6 (25.0%)	13 (54.2%)	5 (20.8%)	8 (33.3%)

Table 4: Session-level categorical outcomes

dimension score of Counter-examples was consistently low across all models (most values were between 1.0 and 2.0), indicating that the models did not tend to ask corrective questions when students gave incorrect answers, which is a core requirement of authentic Socratic teaching (Delic and Bećirović, 2016). One possible explanation for the poor performance of the three models is that they do not produce the dialog-based step-by-step reasoning that judges associated with the Socratic Method. In a Socratic classroom, after posing thought-provoking questions to students, the teacher may review and reflect on their answers to determine what guiding questions to ask next. In this context, it is required for Socratic instructors to maintain the dialogue contexts and engage in adaptive questioning conditioned on the student's improvements and space for improvement (Delic and Bećirović, 2016; Kargupta et al., 2024). However, the three models did not seem to do that. Their low-to-medium scores in "reliance on questions" and "well-formulated questions" show that they have not been focusing on the judges' needs when asking questions. Besides, the judges' experience indicates that they could not solve a large proportion of the problems with the three models' guidance, implying that these models might be giving useless instructions. This finding implies that the three models in this study may not qualify as Socratic tutors in middle school mathematics. In resource-constrained countries or educational environments that rely more heavily on lightweight language models which require less context, these three models may not be suitable for deployment without fine-tuning (Isotani et al., 2023). If these models must be used to support Socratic-style classrooms, educators should exercise caution and carefully review each response generated by the model to confirm that it actually helps students think through the problem.

5 Conclusion and Future Work

This study examined whether three small language models could implement the Socratic Method in a 6th–8th-grade mathematics classroom under controlled prompting conditions, using a rubric-

based evaluation protocol adapted from Souza et al. (2026). We found that all SLMs performed lower than expected when implementing the Socratic Method, with average scores below the median of 3 on a five-point Likert scale. This study highlights the importance of assessing whether the models can help students think critically by generating guiding questions, which does not guaranty effective guidance (Lelièvre et al., 2025; Jeschke et al., 2021).

However, this study has limitations. The scope of this study is limited as only three locally deployed SLMs were examined and all interactions focused on grade 6-8 decimal mathematics tasks, which means that the findings may not be generalized to larger language models, other small language models, other instructional methods and disciplines, and other age groups. Besides, interrater reliability (IRR) statistics were not calculated, due to the dynamics and uncertainty during direct conversations with the SLMs, making the raters code different answers. Therefore, calculating IRR might not be possible considering that it requires multiple raters to assess the same set of targets independently (Shrout and Fleiss, 1979; Koo and Li, 2016). Future works are needed to explore whether the steps from cleaning databases to fine-tuning the models let them fail in instructional methods based on open-ended questions, given that the training data, instruction-tuning procedures, and alignment methods of the three models are not fully transparent. Also, considering that the training corpora and alignment procedures are not fully comparable across models, future work might include comparative research across different SLMs on instructional methods based on open-ended questions to eliminate possible biases.

References

- Benjamin S. Bloom. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay Company, New York.
- Haris Delic and Senad Bećirović. 2016. [Socratic method as an approach to teaching](#). *European Researcher*, 111:511–517.

- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [LLM.int8\(\): 8-bit Matrix Multiplication for Transformers at Scale](#). *Preprint*, arXiv:2208.07339.
- Element Labs Inc. 2025. LM Studio. <https://lmstudio.ai>. Computer software.
- Seiji Isotani, Ig Bittencourt, Geiser Chalco, Diego Dermeval, and Rafael Mello. 2023. *AIED Unplugged: Leapfrogging the Digital Divide to Reach the Underserved*, pages 772–779.
- Colin Jeschke, Christiane Kuhn, Aiso Heinze, Olga Zlatkin-Troitschanskaia, Hannes Saas, and Anke Lindmeier. 2021. [Teachers’ ability to apply their subject-specific knowledge in instructional settings—a qualitative comparative study in the subjects mathematics and economics](#). *Frontiers in Education*, 6:683962.
- Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani-Tur, and Jiawei Han. 2024. [Instruct, not assist: Llm-based multi-turn planning and hierarchical questioning for socratic code debugging](#).
- Terry K. Koo and Mae Y. Li. 2016. [A guideline of selecting and reporting intraclass correlation coefficients for reliability research](#). *Journal of Chiropractic Medicine*, 15(2):155–163.
- Maxime Lelièvre, Amy Waldock, Meng Liu, Natalia Valdes Aspillaga, Alasdair Mackintosh, María Portelo, Jared Lee, Paul Atherton, Robin Ince, and Oliver Garrod. 2025. [Benchmarking the pedagogical knowledge of large language models](#).
- Mateus Monteiro, Aristoteles Barros, Luiz Rodrigues, Diego Dermeval, Ig Ibert Bittencourt, Seiji Isotani, and Valmir Macario. 2025. [“small device, big decision:” comparing lightweight llms’ computational performance and output quality for aied unplugged](#). In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, volume 2592 of *Communications in Computer and Information Science*, pages 160–167. Springer, Cham.
- P. E. Shrout and J. L. Fleiss. 1979. [Intraclass correlations: Uses in assessing rater reliability](#). *Psychological Bulletin*, 86(2):420–428.
- Talita de Paula Cypriano de Souza, Shruti Mehta, Matheus Arataque Uema, Luciano Bernardes de Paula, and Seiji Isotani. 2026. [Can ai be a teaching partner? evaluating chatgpt, gemini, and deepseek across three teaching strategies](#). *Preprint*, arXiv:2603.26673.