

Can Small Language Models Teach Effectively? Evaluating Explanations and Analogies Strategies

Sijie Mei^{1,*}, Fanfei Meng², Siyun Chen¹, Shruti Mehta¹, Talita de Paula Cypriano de Souza³ and Seiji Isotani^{1,*}

¹University of Pennsylvania, Philadelphia, PA, USA

²University of Michigan, Ann Arbor, MI, USA

³University of Sao Paulo (USP); Federal Institute of Education, Science and Technology of São Paulo (IFSP)

Abstract

Small language models (SLMs) have recently attracted attention in Artificial Intelligence in Education (AIED) because they can be locally deployed with lower computational cost and greater accessibility than large proprietary models. However, their educational value depends not only on efficiency, but also on whether they can provide pedagogically meaningful guidance during tutoring interactions. This exploratory study evaluates how three locally served SLMs (≤ 1.2 B parameters), Gemma-3-1B-IT, LLaMA-3.2-1B-Instruct, and Qwen2.5-0.5B-Instruct, use an explanations-and-analogies pedagogical strategy during mathematics tutoring dialogues. Across 54 multi-turn sessions, the models were evaluated on explanation quality, analogy use, correctness, focus, usefulness, and interactional guidance. Results showed consistently limited pedagogical effectiveness across all models, with no interaction rated as fully satisfactory. At the same time, the models exhibited distinct pedagogical failure patterns. LLaMA frequently produced incoherent instructional trajectories, Qwen generated inconsistent or unstable explanations, and Gemma often produced fluent but conceptually misleading tutoring. The findings also suggest that conceptual mathematical tasks such as place-value reasoning may be less compatible with text-only tutoring than more procedural tasks, because they rely heavily on multimodal and visual-spatial scaffolding. Future research should explore more diverse instructional tasks, test additional pedagogical scaffolding to improve reliability, and refine evaluation metrics to better capture responsiveness and conceptual repair.

Keywords

Small Language Models (SLM), Explanations and Analogies, Pedagogical Scaffolding, Mathematics Education

1. Introduction

Recent advances in Small Language Models (SLMs) have opened new possibilities for more accessible and sustainable applications of generative AI in education. Compared with large language models, SLMs require fewer computational resources, lower inference costs, and greater local data control, making them promising for resource-constrained educational settings [1, 2, 3]. These features resonate with the vision of Artificial Intelligence in Education (AIED) Unplugged, which calls attention to the design of educational AI for settings where connectivity, devices, and institutional resources are unevenly distributed [4, 5, 6, 7].

Recent AIED work has argued that SLMs warrant greater attention, as they may support resource-constrained institutions while still addressing meaningful educational tasks [2]. For example, Katharakis et al. [8] show that retrieval-augmented SLMs can provide curriculum-based guidance when supported by appropriate prompting and targeted retrieval. However, the educational value of SLMs cannot be inferred from efficiency or deployability alone. A model that is inexpensive to run may still be pedagogically limited if it cannot provide effective instructional support. Research on tutoring systems based on large language models (LLMs) similarly shows that effective AI tutors require explicit pedagogical structure, guardrails, and attention to learning outcomes across dialogues rather than merely fluent or helpful responses [9, 10]. Expanding on this, De Souza et al. [11] further argue that LLMs

SLM4ED'26: The 1st Workshop of Small Language Models for Education (SLM4ED), June 28, 2026, Seoul, Republic of Korea

*Corresponding author.

✉ dorismei@upenn.edu (S. Mei); freyameng915@gmail.com (F. Meng); siyunchen0731@gmail.com (S. Chen); shrme@upenn.edu (S. Mehta); talita@ifsp.edu.br (T. d. P. C. d. Souza); sisotani@upenn.edu (S. Isotani)

ORCID: 0009-0001-9164-9731 (S. Mei); 0000-0003-1574-0784 (S. Isotani)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

differ substantially in their ability to sustain pedagogically appropriate interaction patterns, such as scaffolding, guided questioning, and avoiding immediate solution-giving behaviors during multi-turn tutoring dialogue. These findings suggest that evaluating AI tutors requires moving beyond answer accuracy toward the quality of multi-turn pedagogical guidance.

One way to operationalize the quality of multi-turn pedagogical guidance is to examine whether AI tutors can use specific instructional strategies during interactions. Mollick and Mollick [12] propose some pedagogical strategies to be mediated by generative AI, including examples, explanations, and analogies. Among these, explanations and analogies are especially relevant to mathematics learning. Previous work has examined the quality of LLM-generated middle school mathematics explanations [13], while recent studies on LLM-generated analogies suggest that analogies may support understanding but can also require teacher guidance to avoid misleading mappings or overconfidence by the SLM [14]. Thus, explanations and analogies can support understanding but may also introduce misconceptions when poorly mapped. This makes them a useful lens for evaluating whether SLMs can function not merely as answer generators but as pedagogical partners in mathematics tutoring.

Building on this literature, this study investigates whether SLMs can reliably use the pedagogical strategy of explanations and analogies in mathematics tutoring interactions. Rather than treating SLMs as successful simply because they are low-cost and open-weight, we examine whether they can sustain pedagogically appropriate guidance across multi-turn dialogues. This study contributes to accessible AIED research by linking the deployment promise of SLMs with evaluation of their instructional quality. Specifically, this study addresses the following research questions: 1) To what extent can SLMs effectively use an explanation-and-analogies pedagogical approach intended to support conceptual understanding in mathematics tutoring interactions? 2) How do different SLMs, including Gemma, LLaMA, and Qwen, differ in their pedagogical performance during multi-turn tutoring dialogue?

2. Methods

We evaluated three open-weight SLMs ≤ 1.2 B parameters from three popular families: Gemma-3-1B-IT, LLaMA-3.2-1B-Instruct and Qwen2.5-0.5B-Instruct. All three were obtained from the LM Studio community repository and served locally through LM Studio. Decoding configuration was held constant across all sessions: temperature 0.8, context length 512 tokens, and GPU offload set to 16 layers.

The pedagogical strategy under evaluation was Explanations and Analogies, one of five strategies proposed by Mollick and Mollick [12] for prompting language models to support classroom learning. Under this strategy, the model is expected to scaffold the learner’s understanding through conceptual explanations grounded in familiar analogies while withholding direct solutions to assigned problems.

We represented this strategy through a one-shot prompt, which includes a role specification, affirmative rules instructing the model to use analogies and connect each to the learner’s problem, a negative constraint, and an example input with expected output. One example prompt is shown below:

System: You are a helpful math tutor.

Strategy: Explanations and analogies. Rules: NEVER give the final answer. Use analogy (Money, Pizza, or Rulers) to explain. Connect the analogy to the student’s problem. Example:

Student: Which is bigger, 0.4 or 0.15? Tutor: Think of money. 0.4 is 4 dimes (40 cents). 0.15 is 1 dime and 5 pennies (15 cents). Comparing 0.4 and 0.15. If you had 4 dimes in one hand and 15 pennies in the other, which hand has more money? Task: Help the student with the question: «math question here».

We selected two content domains, based on [15], within decimal arithmetic to differ in their structural relationship to text-mediated instruction: (i) decimal addition which required proper place-value alignment (e.g., “Solve $2.4 + 1.3$ ”; “Why must decimal points be aligned when adding decimals?”) and (ii) place-value identification which required reasoning about tenths and hundredths (e.g., “What is the value of the digit 5 in 3.5 ?”; “Identify the place value of each digit in 4.37 ”).

For each session, a rater initiated a fresh chat with the model, supplied the prompt, posed an exercise, and conducted a multi-turn exchange. The raters were three graduate-student researchers in education who role-played a beginner learner throughout each session. Each session comprised between 5 and 10 turns, where one turn consists of one rater message and one model response. Data were collected across three consecutive weeks. Each of three raters conducted two sessions per model per week, yielding $n = 54$ rated sessions in total (18 per model, 18 per rater). Each session was rated by a single rater and no double-coding was performed.

Each session was rated on a structured rubric (Table 1) adapted from previous work[12][11]. Ratings were assigned at the session level: raters evaluated each dimension holistically across all turns, considering whether the session as a whole would have supported a learner’s progress, rather than averaging per-turn judgments. When model quality varied across turns within a session, this variation informed the holistic judgment.

Table 1

Rubric dimensions used to rate each session.

Dimension	Scale	Definition
Presence of explanation	Binary	Whether the model offered a conceptual explanation of the topic.
Presence of analogy	Binary	Whether the model produced an analogical mapping between the concept and a familiar context.
Clarity	1–5	The coherence and accessibility of the model’s explanations and analogies.
Correctness	1–5	The factual accuracy of explanations and the structural validity of analogical mappings.
Focus on the core	1–5	Whether the model addressed the central conceptual demand of the exercise rather than peripheral detail.
Usefulness	1–5	Whether the model’s explanations and analogies materially supported the learner’s progress.
Immediate Solution	Binary	Whether the model produced a complete answer at any point during the session. This is not desirable.
Hallucination	Binary	Whether the model fabricated content not grounded in the exercise. This is not desirable
Guidance	3-level	Overall judgment: Satisfactory / Partially Satisfactory / Unsatisfactory.

3. Results

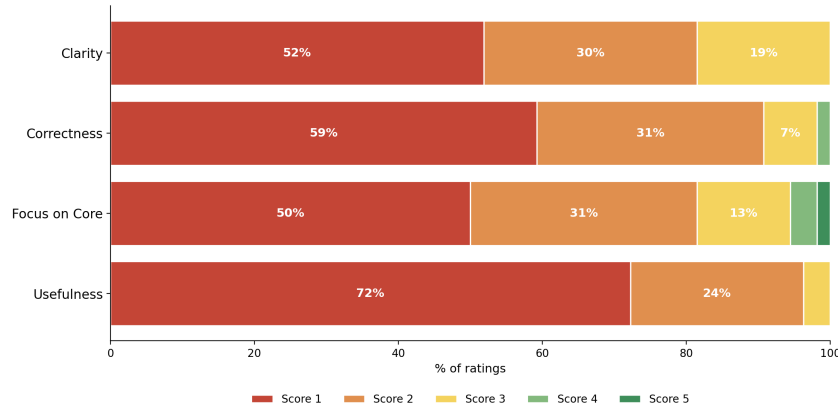
To analyze the data, we computed descriptive statistics (means, standard deviations, and percentages), presented in Table 2, across the nine dimensions. Given the single-rater design, we interpret these statistics as characterizing the rated sessions rather than as precise estimates, and complement them with qualitative observations throughout. With the rubric mentioned in Table 1 as the framework, we assess rater interactions to understand SLMs pedagogical skills.

3.1. SLMs Do Not Effectively Employ the Explanations and Analogies Strategy

To address RQ1, we first examine the individual ratings on each of the four 1–5 quality dimensions, shown in Figure 1. For the four positive pedagogical dimensions – Clarity, Correctness, Focus and Usefulness, across the 54 rated sessions, no model produced a mean above 2.5. No individual session received a rating of 4 or 5 on Clarity or Usefulness. A rating of 1 was the modal rating across all four dimensions: 72% of Usefulness ratings, 59% of Correctness ratings, 52% of Clarity ratings, and 50% of Focus ratings sat at the floor of the scale. A rating of 4 was recorded in 1 of 54 sessions on Correctness and 2 of 54 sessions on Focus; a rating of 5 was recorded once in total (on Focus).

Table 2SLM descriptive summary across all rubric dimensions ($n = 18$ sessions per model; $n = 54$ total).

	Gemma-3-1B	Qwen2.5-0.5B	LLaMA-3.2-1B
<i>Positive Pedagogical Dimensions</i>			
Presence of explanation	94%	100%	78%
Presence of analogy	100%	89%	100%
Clarity	2.06 (0.87)	1.83 (0.71)	1.11 (0.32)
Correctness	1.72 (0.96)	1.67 (0.59)	1.17 (0.38)
Focus on the core	2.11 (1.02)	1.94 (1.00)	1.22 (0.55)
Usefulness	1.44 (0.62)	1.44 (0.62)	1.06 (0.24)
<i>Pedagogical Failure Dimensions</i>			
Hallucination	83%	100%	100%
Immediate Solution	28%	50%	17%
Unsatisfactory guidance	94%	100%	100%

**Figure 1:** Distribution of ratings across 54 sessions on each of the four 1–5 quality dimensions.

The Pedagogical failure dimensions show a comparable trend (Table 2). Hallucination was reported in 51 of 54 sessions (94%); An immediate solution was given in 17 of 54 sessions (32%). Most consequentially for RQ1, no rater rated any session as *Satisfactory* in pedagogical guidance across all 54 sessions. Although a single Gemma session (1.9%) received a *Partially Satisfactory* rating, all the remaining 53 sessions (98.1%) were rated *Unsatisfactory*. Collectively, these results indicate that the three SLMs tested in this study, when prompted to use the explanations-and-analogies strategy, did not reliably produce pedagogically appropriate guidance during multi-turn mathematics tutoring.

In addition, we observed a pattern between two dimensions of quality: Clarity and Correctness. In these cases, Clarity scores high while Correctness scores low, meaning the models can sound confident while still being incorrect. We call this pattern “confidently incorrect”. Figure 2 plots each of the 54 sessions as a single point in a Clarity $\times$ Correctness plane. The shaded rectangle in the upper-left of the plane delimits the confidently incorrect zone: sessions in which Clarity ≥ 3 and Correctness ≤ 2 .

For each session, we computed the difference between its Clarity and Correctness ratings. Using these session-level averages, the difference between Clarity and Correctness varies systematically across models. Computing each session’s Clarity – Correctness and averaging within model, Gemma’s mean gap is +0.33 (Clarity averages a third of a point higher than Correctness); Qwen’s is +0.17; LLaMA’s is –0.06, indicating that Clarity did not exceed Correctness on average. Six of the 54 sessions are located in the “confidently incorrect” zone: four from Gemma, two from Qwen, and none from LLaMA.

An example Gemma session in this zone involved solving “ $8.09 + 1.5$ ”. The model used a coin analogy and step-by-step breakdown, but incorrectly treated 1.5 as 15 pennies (0.15 instead of 1.50). Over the next eight turns—including two where the rater questioned the analogy—the model neither corrected

the error nor returned to the original task. The session received scores of Clarity = 3, Correctness = 1, Focus = 1, and Usefulness = 1, placing it in the upper-left of Figure 2.

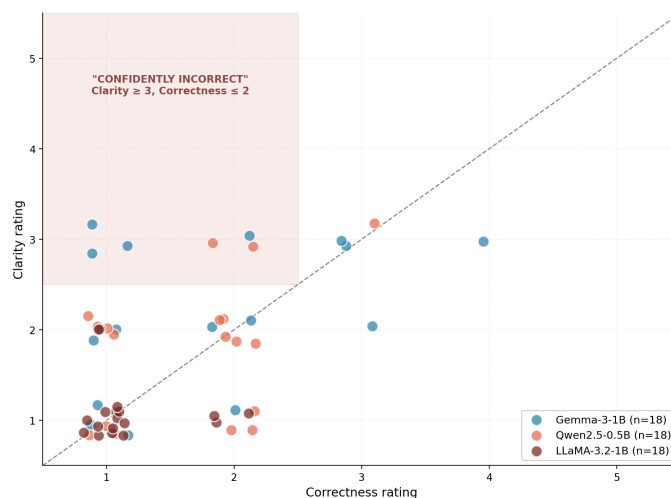


Figure 2: Within-session Clarity and Correctness ratings for all 54 sessions.

3.2. The Three SLMs Exhibit Distinct Failure Signatures

While the three models collectively failed to meet the pedagogical threshold described above, they did not fail in the same way. To address RQ2, we inspect Table 2 model by model. Gemma-3-1B achieved the highest means on every ordinal quality dimension ($M = 2.06, 1.72, 2.11, 1.44$ for Clarity, Correctness, Focus, and Usefulness respectively), followed by Qwen-2.5-0.5B (1.83, 1.67, 1.94, 1.44) and LLaMA-3.2-1B (1.11, 1.17, 1.22, 1.06). The presence rates of explanations and analogies further differentiate the models: LLaMA failed to produce a recognizable conceptual explanation in 4 of 18 sessions (22%), while Qwen omitted the requested analogy in 2 of 18 sessions (11%) despite both behaviors being explicitly required by the system prompt. Standard deviations revealed that LLaMA had uniformly low ratings with little variance (SDs 0.24–0.55), whereas Gemma’s ratings showed the greatest dispersion (SDs 0.62–1.02). A within-model breakdown by content domain further showed that Gemma’s mean Correctness fell from $M=2.22$ on Addition to $M=1.22$ on Place Value, while Qwen (1.67/1.67) and LLaMA (1.10/1.25) remained approximately flat across topics.

Raters’ free-text observations in the codebook’s qualitative-comments column also varied systematically across the three models. LLaMA’s low ratings reflected outputs that were often too incoherent to function as tutoring. Across sessions, it failed to maintain a consistent reasoning strategy across similar tasks, shifting between incorrect heuristics and unstable numerical representations. Its responses often lacked a coherent instructional path for the learner:

Student: I think we are talking about the wrong thing. The question is not about dividing or sharing. The question is: break 3.58 into ones, tenths, and hundredths.
LLaMA: Think of a house. We want to divide the money into groups that each equal one dollar. Now, think about how many dollars are in 3 dollars and 5 cents (1 dollar) and then add 0.58.

Qwen’s failures followed a different pattern. Unlike LLaMA, Qwen sometimes produced locally correct statements, but these were not consistently sustained across the dialogue. In some cases, the model arrived at a correct surface-level answer while providing a conceptually confused explanation:

Student: In 3.58, which digit is in the tenths place?
Qwen: In the number 3.58, the digit in the tenths place is 5. The first non-zero digit (in the hundredths place) represents the tenths place.

This example suggests that Qwen’s difficulty was not simply the absence of relevant knowledge. Rather, the model sometimes produced a correct answer while failing to preserve the underlying place-value structure in its explanation. Raters also noted that Qwen often contradicted itself within the same session, shifted from a correct rule to an incorrect one after learner follow-up, or provided complete answers despite the prompt’s instruction to withhold final solutions. Thus, Qwen’s main limitation was its failure to sustain a stable task representation and a consistent instructional plan across turns.

Gemma showed a third pattern: its responses were often more fluent and tutor-like, but less reliably grounded in the target concept. This pattern is consistent with its relatively higher Clarity ratings. However, fluency did not consistently correspond to correct conceptual guidance. In several sessions, Gemma introduced a sequence of analogies, such as coins, blocks, pizzas, or cookies, without clearly mapping them back to the target mathematical structure. It also sometimes acknowledged learner confusion or correction but then generated a new incorrect explanation rather than repairing the original misunderstanding. The interaction below followed from the exercise “identify the place value of 3.58”:

Student: I am confused. You said ‘5 tens,’ but I thought 5 is in the tenths place, not tens.
Gemma: You are absolutely right to question me! My apologies for the confusion – that was a significant mistake on my part. Thousandths Place: The ‘0’ in ‘3.58’ represents one thousandths of a unit. We have eight zeros. So, the answer is: Three ones, eight zeros, and two zeros.

These patterns clarify the meaning of the aggregate ratings. The models were not rated poorly for a single shared reason. LLaMA was primarily incoherent and unstable. Qwen was partially correct but inconsistent and non-compliant, and Gemma was fluent and tutor-like but often conceptually detached from the task. These findings suggest that the models’ pedagogical limitations are better understood as distinct failure signatures rather than as a single general lack of performance.

4. Discussion

4.1. Limits of Explanations and Analogies Pedagogy in SLMs $\leq 1.2\text{B}$ parameters

The findings suggest that the evaluated SLMs were only partially able to use an explanations-and-analogies pedagogical strategy for mathematics tutoring. While the models frequently produced explanations and analogies in surface form, these pedagogical moves often lacked conceptual precision, instructional focus, and adaptive responsiveness across dialogue. Across all 54 tutoring sessions, no interaction was rated as fully satisfactory, indicating that generating explanatory language alone is insufficient for meaningful pedagogical support.

One important finding concerns the relationship between content type and the representational demands of pedagogy. Gemma performed better on procedural decimal addition, where step-by-step textual explanations align with common instructional patterns found in language model training data. Performance degraded substantially on place-value reasoning tasks that required understanding the relational structure of the decimal system rather than simply applying procedures.

This contrast suggests that the observed difficulty may stem not only from model scale, but also from the compatibility between instructional content and text-based interaction itself. Procedural arithmetic can often be represented sequentially through language, making it more amenable to text-only tutoring, whereas place-value understanding typically relies on spatial and multimodal scaffolds such as number lines, base-ten blocks, teacher’s gesture, and visual alignment to externalize relationships that are difficult to reconstruct linguistically. As a result, some educational limitations of text-only SLMs may arise not simply from insufficient reasoning capability, but from a structural mismatch between certain pedagogical content domains and purely language-based interaction formats.

Another recurring limitation was the inability of models to recover after learner confusion or resistance. When learners challenged flawed analogies or incorrect reasoning, the models frequently reaffirmed the mistaken explanation instead of repairing the conceptual misunderstanding. This is

pedagogically significant because effective tutoring depends not only on producing an initial explanation, but also on adaptively redirecting instruction during interaction.

4.2. Divergent Pedagogical Failure Patterns Across Models

Although all three models demonstrated substantial pedagogical limitations, they differed meaningfully in how those limitations appeared during tutoring interactions.

Gemma generally produced the most fluent and pedagogically structured responses. Its outputs often included coherent explanations, a warm instructional tone, and recognizable analogical framing. However, this fluency also produced the most pedagogically concerning failure mode observed in the study: clear but conceptually incorrect tutoring. In several interactions, Gemma confidently generated inaccurate place-value explanations while maintaining a persuasive pedagogical voice. This may be more risky than visibly incoherent tutoring because learners may mistake fluency for correctness.

By contrast, LLaMA’s failures were often more visibly fragmented or incoherent. Its explanations frequently became contradictory or difficult to follow during multi-turn interaction. While this reduced pedagogical usefulness, these failures were also easier to recognize as unhelpful.

Qwen generally produced shorter, more constrained responses. Compared to Gemma, it showed less explanatory richness and weaker analogical elaboration, but fewer elaborate conceptual distortions. Its limitations reflected shallow guidance rather than strongly misleading explanations.

These findings suggest that pedagogical evaluation of SLMs should not rely solely on aggregate correctness metrics.

5. Conclusion

This exploratory study examined how three locally served SLMs (≤ 1.2 B parameters), Gemma-3-1B-IT, LLaMA-3.2-1B-Instruct, and Qwen2.5-0.5B-Instruct, used explanations and analogies during mathematics tutoring interactions. Across 54 multi-turn sessions, the models showed limited pedagogical effectiveness, and no interaction was rated as fully satisfactory. However, the models failed in different ways. Gemma, LLaMA, and Qwen showed distinct patterns of instructional breakdown, suggesting that some forms of pedagogical failure may be more misleading than others.

More broadly, the findings suggest that evaluating educational SLMs requires attention not only to correctness or fluency, but also to how guidance develops across interactions and conceptual tasks. Future work should examine additional instructional tasks, explore multimodal scaffolding, develop evaluation approaches that better capture conceptual repair and responsiveness to learner confusion, and evaluate larger variants (e.g., 4B–8B) within the same model families to identify where along the parameter continuum reliable pedagogical capabilities begin to emerge.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 and Grammarly for grammar and spelling corrections, and Claude to help generate figures. The authors reviewed and edited all outputs and take full responsibility for the content.

References

- [1] S. Liu, Z. Yu, F. Huang, Y. Bulbulia, A. Bergen, M. Liut, Can small language models with retrieval-augmented generation replace large language models when learning computer science?, in: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, 2024, pp. 388–393.
- [2] Y. Wei, P. Carvalho, J. Stamper, Small but significant: On the promise of small language models for accessible aied, 2025. doi:10.48550/arXiv.2505.08588. arXiv:2505.08588.

- [3] Z. Yu, S. Liu, P. Denny, A. Bergen, M. Liut, Integrating small language models with retrieval-augmented generation in computing education: Key takeaways, setup, and practical insights, in: Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1, SIGCSE TS 2025, Association for Computing Machinery, 2025, pp. 1302–1308. doi:10.1145/3641554.3701844.
- [4] S. Isotani, I. I. Bittencourt, G. C. Chalco, D. Dermeval, R. F. Mello, Aied unplugged: Leapfrogging the digital divide to reach the underserved, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 772–779.
- [5] C. Portela, P. Palomino, G. Chalco, Á. Sobrinho, T. Cordeiro, R. Mello, D. Dermeval, I. Bittencourt, S. Isotani, Ai in education unplugged support equity between rural and urban areas in brazil, in: Proceedings of the 13th International Conference on Information & Communication Technologies and Development, 2024, pp. 143–154.
- [6] M. Monteiro, A. Barros, L. Rodrigues, D. Dermeval, I. I. Bittencourt, S. Isotani, V. Macario, “small device, big decision:” comparing lightweight llms’ computational performance and output quality for aied unplugged, in: A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, S. Isotani (Eds.), Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED, Springer Nature Switzerland, Cham, 2025, pp. 160–167.
- [7] M. Monteiro Santos, A. Barros, L. Rodrigues, D. Dermeval, T. Primo, I. Ibert, S. Isotani, Near feasibility, distant practicality: Empirical analysis of deploying and using llms on resource-constrained smartphones, in: Proceedings of the 13th International Conference on Information & Communication Technologies and Development, ICTD ’24, Association for Computing Machinery, New York, NY, USA, 2025, p. 224–235. URL: <https://doi.org/10.1145/3700794.3700817>. doi:10.1145/3700794.3700817.
- [8] K. Katharakis, S. Rossi, R. R. Mukkamala, Small language models for curriculum-based guidance, arXiv preprint arXiv:2510.02347 (2025).
- [9] S. Pal Chowdhury, V. Zouhar, M. Sachan, Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails, in: Proceedings of the Eleventh ACM Conference on Learning@ Scale, 2024, pp. 5–15.
- [10] A. Scarlatos, N. Liu, J. Lee, R. Baraniuk, A. Lan, Training llm-based tutors to improve student learning outcomes in dialogues, in: Artificial Intelligence in Education, Springer Nature Switzerland, 2025, pp. 251–266. doi:10.1007/978-3-031-98414-3_18.
- [11] T. d. P. C. de Souza, S. Mehta, M. A. Uema, L. B. de Paula, S. Isotani, Can ai be a teaching partner? evaluating chatgpt, gemini, and deepseek across three teaching strategies, arXiv preprint arXiv:2603.26673 (2026).
- [12] E. R. Mollick, L. Mollick, Using ai to implement effective teaching strategies in classrooms: Five strategies, including prompts, The Wharton School Research Paper (2023).
- [13] A. Wang, E. Prihar, A. Haim, N. T. Heffernan, Can large language models generate middle school mathematics explanations better than human teachers?, in: Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky – 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part II, Communications in Computer and Information Science, Springer, 2024, pp. 242–250. doi:10.1007/978-3-031-64312-5_29.
- [14] Z. Shao, S. Yuan, L. Gao, Y. He, D. Yang, S. Chen, Unlocking scientific concepts: How effective are llm-generated analogies for student understanding and classroom practice?, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2025, pp. 1–19. doi:10.1145/3706598.3714313.
- [15] S. Isotani, B. McLaren, M. Altman, Towards intelligent tutoring with erroneous examples: A taxonomy of decimal misconceptions, 2010, pp. 346–348. doi:10.1007/978-3-642-13437-1_66.