ArguMath: AI-Simulated Environment for Pre-Service Teacher Training in Orchestrating Classroom Mathematics Argumentation

Jiwon Chun, Yuling Zhuang, Armanto Sutedjo, Colin Xu, Rong Ren, and Meng Xia

Texas A&M University, College Station, TX 77843, USA
{jiwonchun,ylzhuang,asutedjo,xucol000,rren,mengxia}@tamu.edu

Abstract. Facilitating productive mathematical argumentation, especially asking rational questions, is essential yet remains challenging for pre-service mathematics teachers (PMTs), who often have limited opportunities to apply abstract theoretical knowledge in authentic practice. At the same time, recent advances in large language models (LLMs) have expanded the potential for simulating students in educational settings, enabling low-risk environments for instructional practice. To inform the design of a system that supports PMTs in orchestrating classroom argumentation, we conducted a formative study with eight experienced mathematics teachers to identify key design requirements, including personalization, realistic simulations, structured reflection, and ease of use. Building on these requirements, we developed *ArguMath*, an AI-simulated classroom environment that supports PMTs in practicing the orchestration of mathematical argumentation. *ArguMath* comprises three core components: (1) customization of classroom settings; (2) simulation of classroom discussions with AI-based students grounded in authentic transcripts and augmented with real-time instructional suggestions; and (3) structured reflection through discourse annotation and overall feedback. Results from an exploratory user study with seven PMTs, complemented by interviews with four experienced teachers, indicate that *ArguMath* has the potential to support PMTs' classroom orchestration skills, particularly theory-aligned questioning strategies.

Keywords: Mathematics teacher education · LLM-based Simulation · Classroom Argumentation

1 Introduction

Mathematical argumentation, a process in which students construct and justify claims with evidence, is fundamental to enhancing mathematical understanding and problem-solving skills [4]. To foster students' productive engagement, teachers must effectively orchestrate discourse by posing purposeful questions that elicit and connect student reasoning [2]. However, pre-service mathematics teachers (PMTs) often lack sufficient opportunities to practice these complex facilitation skills due to limited class time and the high demands of individualized

instructor feedback [11]. Current preparation relies heavily on passive observation, which is often insufficient to bridge the gap between theoretical knowledge and classroom enactment [1,12]. Therefore, there is a pressing need for low-risk, authentic training environments where PMTs can actively rehearse instructional strategies to prepare for the complexities of real classroom contexts [10].

LLMs have transformed classroom simulations, offering risk-free environments for teaching innovation [15,6]. While applications facilitate engaging teacher training, they primarily simulate small student groups, often resulting in idealized or misaligned interactions with the nuances of real classroom practice [8,14]. Furthermore, few studies focus on the orchestration of mathematical argumentation as a core instructional practice for PMTs. Prior research has largely emphasized general performance feedback rather than integrating established pedagogical frameworks into real-time guidance and reflection [12].

To address these gaps, we introduce *ArguMath*, an AI-based classroom simulator designed to scaffold PMTs' argumentation orchestration skills. Informed by a formative study with eight in-service mathematics teachers on the practice of argumentation for PMTs, *ArguMath* has three components: (1) a personalized interface to set parameters including grade level, topic, and student knowledge; (2) authentic argumentation simulations powered by Retrieval-Augmented Generation (RAG) using TIMSS Video[1]; and (3) pedagogical integration of frameworks (Toulmin's argumentation model [9], Teacher Rational Questioning Framework (TRQF) [16]) to provide real-time guidance and structured reflection.

To evaluate *ArguMath*, we compared it against a baseline system that reflects traditional PMT preparation, in which participants review scenarios and receive generic, non-interactive feedback. An exploratory user study with seven PMTs and four experienced teachers was conducted to address two research questions. **RQ1**: How does PMTs' argumentation orchestration, specifically questioning performance, differ between *ArguMath* and the baseline? **RQ2**: How do PMTs and experienced teachers perceive *ArguMath* compared to the baseline system? Results have suggested evidence that *ArguMath* better supports PMTs' classroom orchestration and theory-aligned questioning strategies. Participants also rated the simulation and reflection features as highly useful and intuitive, emphasizing their value for professional development.

2 Formative Study and Design Requirements

To inform the system design for PMTs' argumentation orchestration, we interviewed eight experienced mathematics teachers familiar with PMT training (6 female, 2 male) with an average of 18.13 years of experience ($SD = 11.29$), covering Grades 6–12 and class sizes of 4–32. Each one-hour Zoom interview covered five areas: current classroom discourse practices, novice teacher challenges, PMT support needs, AI experiences, and feedback on a early-stage prototype which is designed a chat interface with multiple AI-simulated students. Participants reviewed the interface and provided suggestions.

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

Based on our formative study, we distilled four key design requirements. **DR1.** Provide a personalized interface to adjust the simulation environment for training, including grade level, math topic, and classroom dynamics. Teachers emphasized that adapting strategies to diverse classroom compositions is a core challenge for PMTs. **DR2.** Deliver realistic AI-based simulations to help PMTs navigate unpredictable classroom dynamics. Participants valued rehearsing with students of varied understanding and engagement levels, as well as requested immediate feedback on questioning clarity and probing depth. **DR3.** Integrate structured and theory-based reflection activities guided by pedagogical frameworks to analyze discourse patterns. Participants suggested that annotating teacher-student interactions helps PMTs identify and refine their orchestration strategies. **DR4.** Ensure an intuitive user experience with visual cues (e.g., emojis) to facilitate quick perception of student engagement. Reducing cognitive burden allows PMTs to focus on the complexity of argumentation.

3 System

Integrating the design requirements, we present *ArguMath* to support PMTs in practicing classroom argumentation facilitation. Leveraging OpenAI's GPT-4o, the system employs prompt engineering to achieve a three-step workflow.

Step 1: Context Personalization PMTs first configure the classroom simulation by specifying three parameters: *Grade Level*, the simulated students' grade; *Math Topic*, the instructional content; and *Class Description*, for which PMTs provide a brief summary of student demographics, engagement, and prior mathematical understanding (e.g., "about half of students are highly engaged, while others struggle with applying algebraic concepts"). These inputs generate a personalized pedagogical context that governs student behaviors and interaction logic in subsequent steps.

Step 2: Classroom Simulation Reflecting the average class size from our formative study ($M = 24$), we simulated a classroom of twenty students (Figure 1). While initial prompt-based generation resulted in formal and unnatural responses (e.g., "It's a convention in algebra"), we transitioned to Retrieval-Augmented Generation (RAG), leveraging authentic classroom transcripts [7]. *ArguMath* generates more natural utterances characterized by human-like hesitation and explanatory reasoning (e.g., "We don't include the multiplication sign because it can look confusing next to a variable, like the letter x").

Dataset Construction To generate a realistic simulation, we utilized publicly available classroom videos and transcripts from TIMSS Video[2]. Twelve eighth-grade mathematics lessons with English transcripts were selected. Two authors segmented the transcripts and annotated them with speaker roles. We then constructed 214 unique student profiles by using an LLM to extract and summarize each student's participation patterns (Teacher call, Voluntary, Mixed), engagement level (Low, Medium, High), mathematical level (Beginner, Beginner-Intermediate, Intermediate, Intermediate-Proficient, Proficient), argumentation

[2] https://www.timssvideo.com/

level (None, Statement only, Simple reasoning, Partial reasoning, Justification, Application reasoning, Reasoning with justification, Guidance reasoning, Clarification), and typical utterances from the transcripts [13]. These profiles were subsequently verified and refined by the two authors.

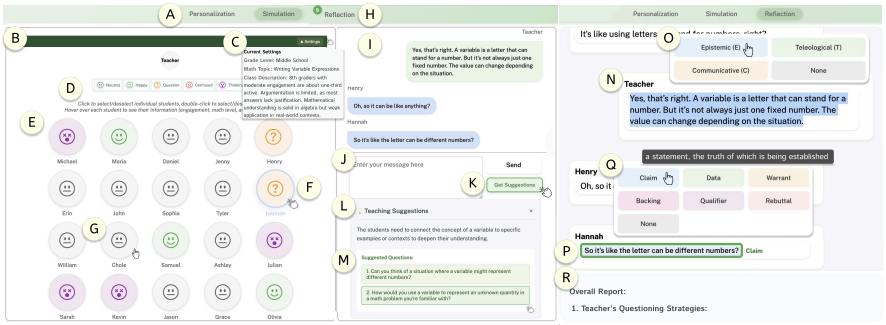

Fig. 1. Interface of *Step 2: Classroom Simulation* (left) and *Step 3: Strategy Reflection* (right). (A) Navigation bar. (B) Virtual classroom with AI-simulated students. (C) *Settings* from *Step 1*. (D) Emoji and interaction legends. (E) Student faces and speech bubbles. (F) Student selection controls. (G) Hover view of student profiles. (H) Responding student count. (I) Full chat view. (J) Text input. (K) *Get Suggestions* for questioning strategies feedback. (L) Teaching strategy suggestions. (M) Recommended questions. (N) Full chat from *Step 2*. (O, Q) Dialogue labeling using TRQF and Toulmin categories. (P) Label confirmation. (R) Overall feedback.

Student Selection The system leverages RAG to select 20 student profiles from a 214 dataset based on *Step 1* settings. Selection criteria include engagement, mathematical proficiency, and argumentation levels. At onset, these profiles are mapped to virtual identities (Figure 1, left).

Student Response and Emoji Generation At the start of the simulation, students remain silent with neutral emojis. Once a teacher poses a question (Figure 1 J), *ArguMath* generates real-time student responses and affective emojis (Figure 1 E, I). If the teacher does not specify respondents, the system selects them based on the established classroom context and student profiles. To ensure authenticity, generated utterances incorporate fillers, contractions, and expressions of confusion or excitement. In parallel, *ArguMath* updates student emojis (neutral, happy, curious, confused, thinking). This allows PMTs to quickly perceive classroom participation.

Suggestion ArguMath provides real-time guidance on questioning strategies by integrating the classroom context and conversation history (Figure 1 L, M). The system evaluates teacher questions using the TRQF and analyzes student responses using Toulmin's model. Integrating these pedagogical frameworks and classroom context, *ArguMath* generates a single-sentence reasoning suggestion. To maintain clarity, the system omits explicit framework references while providing two recommended questions for the next turn.

Step 3: Strategy Reflection After the simulation, PMTs annotate the dialogue to analyze their facilitation strategies. They then access overall feedback and improvement suggestions (Figure 1, right).

Annotation and Overall Feedback Using an integrated interface [3], PMTs label teacher questions based on the TRQF (*Epistemic*, *Teleological*, *Communicative*) and student responses using Toulmin's elements (*Claim*, *Data*, *Warrant*, *Backing*, *Qualifier*, *Rebuttal*) (Figure 1 O, Q). The system provides immediate verification of these labels; a domain expert evaluation of two 20-turn dialogues confirmed an annotation accuracy of approximately 95%. After PMTs submit a brief self-reflection, the system provides suggestions for improvement and supports follow-up inquiries (Figure 1 R).

4 Evaluation

To address RQ1 and RQ2, we conducted an exploratory user study with seven PMTs and interviews with four experienced mathematics teachers.

4.1 User Study

Participants We recruited seven female PMTs (M_{age} = 21, U1–U7) from an undergraduate teacher education program. Participants reported AI usage (3 often, 2 sometimes, 2 rarely), general comfort with new technologies (5/7 positive). Their teaching experience averaged 0.64 years ($SD = 0.48$).

Baseline We created a baseline system which follows standard PMT training by focusing on the analysis of classroom transcripts. It provides theoretical guidance on the TRQF and Toulmin's model, followed by text-only transcripts from the TIMSS Video (e.g., *Writing Variable Expressions* or *Ratios and Division*). Participants analyze these transcripts using the provided frameworks through an input field, with correct answers displayed only after submission.

Procedure Following a within-subjects design, participants completed two training cycles using the baseline and *ArguMath* in counterbalanced order [8]. Each session was followed by post-task surveys; while the baseline survey focused on usability, the *ArguMath* survey included step-specific evaluations. All testing sessions utilized *ArguMath* with the *Suggestion* feature disabled to evaluate independent facilitation skills. The study was controlled at the middle school level (Grades 6–8) with assigned topics (e.g., *Writing Variable Expressions*, *Ratios and Division*) and class descriptions. Training contexts featured moderate engagement and strong algebraic foundations. In contrast, testing contexts featured lower engagement with diverse mathematical proficiency levels. The simulation starts with a math problem to solve, and participants were instructed to prioritize facilitating mathematical argumentation over task completion.

Testing Performance Evaluation To compare system effectiveness, we analyzed testing dialogues and student prompting frequencies for all participants (U1–U7) across both conditions. Two researchers independently coded teacher questions and student responses. Discrepancies were resolved through iterative discussion until a consensus was reached.

4.2 Experienced Teacher Interview

We conducted interviews with experienced mathematics teachers to evaluate *ArguMath*'s workflow, feature utility, and potential for pedagogical improvement. Four experienced female mathematics teachers were recruited and their teaching experience averaged 12.25 years ($SD = 4.03$), covering middle (Grades 6–7) and high school (Grades 9–12) with class sizes of 4–30 students. While three experts (E1–E3) participated in the formative study, E4 was newly recruited. Following a system introduction, experts explored *ArguMath* using self-selected parameters for grade level, math topic, and class description in *Step 1*. After the session, participants completed a questionnaire evaluating the overall workflow, specific features, and potential areas for future improvement.

4.3 Results

To address RQ1 and RQ2, we report quantitative analyses of interaction logs and post-task ratings, followed by qualitative findings from expert interviews.

Table 1. Performance of seven participants (U1–U7) comparing the baseline system (left) and *ArguMath* (right) in terms of elicited student responses, teacher questions (TRQF categories), and student responses (Toulmin elements).

	Baseline											***ArguMath***										
Id	**N**	**TRQF**			**Toulmin**							**N**	**TRQF**			**Toulmin**						
		E	T	C	Cl	Da	Wa	Ba	Qu	Re			E	T	C	Cl	Da	Wa	Ba	Qu	Re	
U1	4	2	-	2	10	-	-	-	-	3		8	1	2	3	12	-	2	-	1	1	
U2	5	2	2	3	8	-	3	2	-	-		8	1	2	1	5	-	3	3	-	1	
U3	20	3	-	1	32	-	4	3	-	-		20	3	1	-	25	-	22	1	-	2	
U4	8	3	7	4	8	2	2	-	1	1		7	9	6	4	21	4	8	2	-	-	
U5	12	2	-	-	10	-	6	-	1	5		20	-	2	-	15	-	20	-	6	1	
U6	20	-	-	-	5	-	27	-	28	-		20	5	4	16	-	-	12	-	38	8	
U7	7	10	1	7	7	-	-	-	7	2		6	1	4	3	-	16	-	-	6	-	
Sum	76	22	10	17	80	2	42	5	37	11		89	20	21	27	78	20	67	6	51	13	

RQ1: Performance Differences Between ArguMath and Baseline Overall, *ArguMath* led to higher quantities and a broader diversity of theory-aligned interactions than the baseline system (Table 1). Quantitative analysis of interaction logs showed that *ArguMath* elicited more student responses (89 vs. 76), TRQF-coded teacher questions (68 vs. 49), and Toulmin-coded student responses (235 vs. 177). These results suggest that *ArguMath* effectively promotes higher engagement and more frequent application of pedagogical frameworks compared to traditional training.

To assess diversity, we calculated normalized Shannon's entropy,

$H = (-\sum_{i=1}^{k} p_i \ln p_i)/\ln k$, yielding an evenness score between 0 and 1. *ArguMath* achieved higher diversity scores than the baseline for both TRQF (0.99 vs. 0.96) and Toulmin (0.85 vs. 0.75) distributions. These results reflect more diverse pedagogical strategies and balanced argumentation. In particular,

ArguMath prompted more Teleological questions to elicit student methodologies, for instance, U2 asked, "Why did you multiply by 0.3 instead of 30 or 3?"

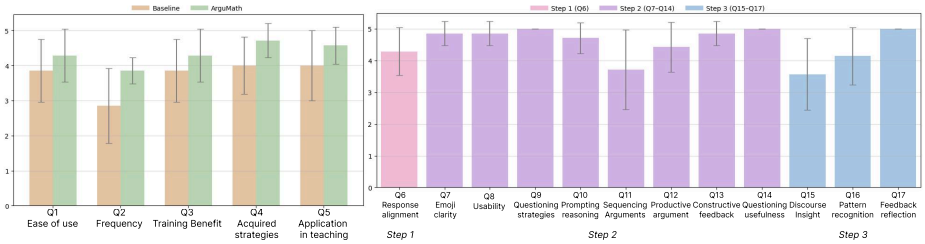

Fig. 2. Means and standard errors (5-point Likert scale) for overall system comparison (Q1–Q5, left) and *ArguMath*-specific feature ratings (Q6–Q17, right).

RQ2: Perceptions of ArguMath and Baseline As shown in Figure 2, *ArguMath* received higher ratings than the baseline across all measures, notably in future use ($M = 3.86\,(0.38)$ vs. $M = 2.86\,(1.07)$), argumentation orchestration practice ($M = 4.29\,(0.76)$ vs. $M = 3.86\,(0.90)$), perceived helpfulness for acquiring strategies ($M = 4.71\,(0.49)$ vs. $M = 4.00\,(0.82)$), and classroom applicability ($M = 4.57$ vs. 4.00).

Step-specific ratings confirm the efficacy of the classroom simulation (*Step 2*) and strategy reflection (*Step 3*). Participants highly rated *ArguMath* for practicing questioning strategies (Q9: $M = 5.00$), prompting students to articulate their reasoning (Q10: $M = 4.71\,(0.49)$), and utilizing suggestions and feedback to guide instructional moves (Q12: $M = 4.43\,(0.79)$; Q13: $M = 4.86\,(0.38)$; Q14: $M = 5.00\,(0.00)$). Qualitative feedback highlighted the system's realism; E1 noted that simulated responses mirrored actual student dialogue, such as "oh, you mean like this." The interface was also perceived as intuitive (Q7–8: $M = 4.86, (0.38)$), with U6 praising the clarity of visual elements. However, lower ratings for sequencing student contributions (Q11: $M = 3.71\,(1.25)$) suggest challenges in managing discourse flow. U1 reported feeling "overwhelmed" when multiple student answers appeared simultaneously. Regarding *Step 3*, features for noticing reasoning patterns (Q16: $M = 4.14\,(0.90)$) and providing actionable feedback (Q17: $M = 5.00\,(0.00)$) were well-received. E2 emphasized that labeling question types helped identify a lack of variety in her pedagogical approach.

5 Discussion

Results suggested that *ArguMath* supports the development of theory-aligned questioning strategies, and we further analyze the design implications as follows.

Student Simulation using Authentic Classroom Discourse Realistic simulations of student struggles are crucial for effective pedagogical role-play.

Participants rated the *ArguMath* simulation as highly realistic, noting that varied engagement, informal utterances, and pauses closely mirrored actual classroom interactions. Such realism motivated participants to iteratively refine their questioning strategies. While simple prompt engineering often lacks student-like hesitation or reasoning [8], *ArguMath* leverages Retrieval-Augmented Generation (RAG) based on 214 student profiles. This approach reflects diverse reasoning quality and participation patterns. Furthermore, unlike prior work limited to small-group chat interfaces [8], our emoji-based design supports classroom-scale argumentation. Participants valued the interface for its intuitiveness.

Structured Theory-based Reflection A key feature of *ArguMath* is its structured reflection scaffold, requiring PMTs to annotate their questions via TRQF and interpret student responses through Toulmin's model. Participants identified these features as the system's most valuable aspects (Figure 2). The structured scaffold shifted PMTs' participation from simply generating questions to active diagnosis of instructional moves. By examining how specific questioning types elicited argument components, PMTs engaged in a rigorous cycle of practice and reflection. Aligning with the "approximations of practice" framework [5], *ArguMath* integrates teaching rehearsals with theoretically grounded analysis. However, effective implementation requires that PMTs are familiarized with the underlying pedagogical theories prior to the annotation activities.

Limitation and Future Work First, this study was conducted in a controlled setting with fixed grades and topics, potentially limiting generalizability. As simulated students were derived from a single dataset, they may not capture broader linguistic diversity. Future research should test *ArguMath* across more diverse classroom dynamics. Second, the text-only interface restricts applicability to visually intensive domains like geometry or functions, where graphs and figures are critical for reasoning. Future iterations should incorporate equation and figure-support features. Third, more sample sizes and longitudinal design are required in future studies to enable more robust statistical analyses.

6 Conclusion

We present *ArguMath*, an LLM-based classroom simulator that supports PMTs in orchestrating mathematical argumentation through context personalization, retrieval-augmented student simulation, and theory-based reflection and feedback. Findings from a within-subjects study and expert interviews suggest that these features help PMTs practice diverse theory-aligned questioning strategies, better understand students' mathematical reasoning, and strengthen PMTs' readiness to orchestrate mathematical argumentation. Together, the results indicate the potential of AI-simulated classrooms to support the integration of theory and practice in mathematical argumentation within teacher preparation.

References

1. Arseven, I.: The use of qualitative case studies as an experiential teaching method in the training of pre-service teachers. International Journal of Higher Education

7(1), 111–125 (2018)
2. Brown, R.: Using collective argumentation to engage students in a primary mathematics classroom. Mathematics Education Research Journal **29**(2), 183–199 (2017)
3. Chun, J., Zhang, G., Xia, M.: Conflictlens: Llm-based conflict resolution training in romantic relationship. In: Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. pp. 1–3 (2025)
4. Conner, A., Singletary, L.M., Smith, R.C., Wagner, P.A., Francisco, R.T.: Teacher support for collective argumentation: A framework for examining how teachers support students' engagement in mathematical activities. Educational Studies in Mathematics **86**(3), 401–429 (2014)
5. Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., Williamson, P.W.: Teaching practice: A cross-professional perspective. Teachers college record **111**(9), 2055–2100 (2009)
6. Jin, H., Yoo, M., Park, J., Lee, Y., Wang, X., Kim, J.: Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–28 (2025)
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
8. Pan, S., Schmucker, R., Garcia Bulle Bueno, B., Llanes, S.A., Albo Alarcón, F., Zhu, H., Teo, A., Xia, M.: Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–18 (2025)
9. Toulmin, S.E.: The uses of argument. Cambridge university press (2003)
10. Wagner, P.A., Smith, R.C., Conner, A., Francisco, R.T., Singletary, L.: Using toulmin's model to develop prospective teachers' conceptions of collective argumentation. North American Chapter of the International Group for the Psychology of Mathematics Education (2013)
11. Wagner, P.A., Smith, R.C., Conner, A., Singletary, L.M., Francisco, R.T.: Using toulmin's model to develop prospective secondary mathematics teachers' conceptions of collective argumentation. Mathematics Teacher Educator **3**(1), 8–26 (2014)
12. Wess, R., Priemer, B.: Pre-service teachers' professional vision of argumentation opportunities in their analysis of videotaped science classroom episodes. Journal of Science Teacher Education **36**(3), 397–423 (2025)
13. Wu, T., Chen, J., Lin, W., Li, M., Zhu, Y., Li, A., Kuang, K., Wu, F.: Embracing imperfection: Simulating students with diverse cognitive levels using llm-based agents. arXiv preprint arXiv:2505.19997 (2025)
14. Xu, S., Wen, H.N., Pan, H., Dominguez, D., Hu, D., Zhang, X.: Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–26 (2025)
15. Xu, S., Zhang, X.: Leveraging generative artificial intelligence to simulate student learning behavior. arXiv preprint arXiv:2310.19206 (2023)
16. Zhuang, Y., Conner, A.: Teachers' use of rational questioning strategies to promote student participation in collective argumentation. Educational Studies in Mathematics **111**(2), 345–365 (2022)