CHARACTERIZING TEACHERS' MATHEMATICAL CONVERSATIONS WITH LLMS THROUGH THE LENS OF MATHEMATICALLY RESPONSIVE TEACHING

Nadav Marco[1], Raz Harel[1], Gil Schwarts[2]

[1]David Yellin Academic College of Education, [2]Weizmann Institute of Science

This exploratory study proposes a novel approach for integrating communication with large language models (LLMs) into teacher education, and a theoretical framework for analyzing the resulting interactions between prospective teachers (PTs) and artificial intelligence (AI) agents. The study involved 20 PTs who were tasked with identifying a mathematical problem from the curriculum where the LLM generates an erroneous solution, and subsequently guiding it to identify and correct its error. To investigate PT–AI interactions, particularly teachers' mathematical responsiveness (or lack thereof), we employed methods from Systemic Functional Linguistics (SFL). We found substantial variation in responsiveness, shaped by whether PTs framed the interaction as a discussion with AI or adopted a teaching–learning stance.

INTRODUCTION

Recent advances in large language models (LLMs) have enabled extended, context-sensitive dialogue between human users and artificial intelligence (AI) systems across a wide range of domains, including mathematics education (Taani & Alabidi, 2024). While LLMs can generate fluent and often plausible mathematical discourse, their outputs are known to include fragile justifications and confidently stated errors, which can lead human interlocutors to accept flawed mathematical reasoning without sufficient critical examination (Rizos & Gkrekas, 2025; Ergene & Caylan-Ergene, 2025). At the same time, these limitations can be reframed as pedagogical affordances: because LLM-generated mathematical reasoning is imperfect, it creates opportunities for teachers to develop critical, disciplined engagement with mathematical claims through critique and reformulation. Marco and Stylianides (2025) argue that LLM errors can be repurposed as pedagogical leverage for genuine mathematical activity, including problem posing, proof validation, and justification repair.

In this paper, we examine the affordances of teacher–LLM mathematical dialogue for teacher learning through the construct of *mathematically responsive teaching*, an approach that foregrounds how teachers attend to and respond to learners' mathematical ideas, perspectives, and experiences in their interactions (see below). From this perspective, responding productively to errors is not a matter of correction alone, but of engaging with the reasoning that gives rise to them and using that reasoning as a resource for further mathematical sense-making.

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 555-562. PME.

Relatedly, recent work on AI-based teaching simulations engages teachers in eliciting mathematical reasoning and addressing errors through dialogue with AI-based student agents (e.g., Schwarts & Czaczkes, submitted; Zhuang & Zhang, 2025). Both teacher–LLM mathematical conversations and AI-based teaching simulations are relatively recent and rapidly evolving. While they share a focus on responsiveness, their conceptual and methodological connections remain emergent. The present study constitutes an initial step toward bridging these approaches by positioning teacher–LLM exchanges as a site for engaging in mathematically responsive interaction.

Specifically, we examine how prospective teachers (PTs) take up and respond to the LLM's mathematical errors, focusing on PTs' conversational scaffolding and dialogic uptake through the lens of mathematically responsive teaching. This baseline analysis is intended to contribute to a broader research agenda aimed at understanding and supporting PTs' development of subject-specific responsiveness. Accordingly, our research question is: What features characterize teachers' mathematical conversations with LLMs when analyzed through the lens of mathematically responsive teaching?

THEORETICAL UNDERPINNINGS

Mathematically Responsive Teaching

Responsive teaching refers to instructional practice in which teachers attend to, interpret, and build from students' thinking as it emerges in interaction (Jacobs et al., 2010; Robertson et al., 2016). In mathematics education, such responsiveness extends beyond evaluating correctness to recognizing the mathematical substance of learners' ideas, probing their justifications, and responding in ways that advance conceptual understanding while preserving learners' agency (Bishop et al., 2022).

Mathematically responsive teaching (MRT), as conceptualized in this study, foregrounds the relational dimensions of such practice. Drawing on Buber's (1958) account of dialogue as an encounter grounded in openness, presence, and mutual recognition, MRT emphasizes the teacher's responsibility to take the learner's mathematical perspective in interaction. From this stance, teaching involves seeking to understand how the mathematics is experienced and reasoned about by the learner, and responding in ways that treat learners' ideas as meaningful contributions rather than merely as objects of evaluation or correction. Students' contributions thus become resources for questioning, elaboration, and the ongoing development of mathematical meaning (Schwarts, 2025). Responsiveness is realized moment by moment as teachers adapt instruction to learners' experiences, sustaining a dialogic space for mathematical work rather than advancing a predetermined script.

Operationalizing responsiveness through Systemic Functional Linguistics

To operationalize mathematically responsive teaching in teacher–LLM exchanges, we draw on Systemic Functional Linguistics (SFL) as an analytic lens. SFL (Halliday & Matthiessen, 2004) conceptualizes language as realizing three metafunctions: ideational (construing experience and disciplinary content), interpersonal (enacting

roles and stances), and textual (organizing discourse). This framework is well-suited to examine MRT, as MRT concerns how teachers attend to others' mathematical thinking in interaction—work that requires coordinating mathematical content, interactional positioning, and discursive focus.

In this paper we focus on selected SFL systems aligned with core features of MRT and with genuine mathematical dialogue (hereafter: MRT-SFL frameowrk). From the interpersonal metafunction, we examine *Mood* and *Modality* as resources for how PTs open or constrain space for LLM reasoning. Mood indexes interactional roles (e.g., dialogic partner vs. epistemic authority), whereas Modality captures how teachers qualify mathematical claims as tentative, negotiable, or warranting further justification. Together, these resources foreground responsiveness in exchanges with a fallible conversational partner. From the textual metafunction, we examine the *Theme* system, albeit in a reduced manner. Specifically, we examine whether the *topical Theme* of PTs' uttarnces following an LLM contribution takes up the substance of the LLM's preceding reasoning, as an indicator of uptake in responsive dialogue (see further operationalization in the Methods section). We focus on the systems of Mood, Modality, and Theme because together they provide a theoretically grounded way to trace how responsiveness is enacted in interaction: Mood and Modality capture how teachers position themselves and the mathematical claims at stake, while Theme captures what is treated as salient for further collective work. This focus enables analysis of dialogic uptake and scaffolding in teacher–LLM exchanges.

METHODS

The present study employs an exploratory qualitative design, supplemented by descriptive quantification, to characterize PTs' mathematically responsive teaching in interactions with LLMs. As an initial step, we examine whether and how PTs engage in genuine mathematical dialogue with LLMs—specifically, whether their interactions exhibit listening, uptake, and responsiveness to the LLM's mathematical reasoning. The study involved 20 PTs enrolled in a college course taught by one of the authors. Informed consent was obtained from all participants, and all data were anonymized prior to analysis.

Study design

Participants interacted with general-purpose LLMs of their choice (e.g., ChatGPT, Claude, Gemini). We used freely available general-purpose LLMs to ensure immediate access without prior training and to establish an exploratory baseline for future development of specialized AI-based teaching simulators. The course instructor first modelled, in class, a responsive dialogue with ChatGPT, presenting a geometric task that elicited an error from the Chat and demonstrating how to guide the Chat toward self-correction through conceptual questioning rather than direct explanation. Following Marco and Stylianides (2025), PTs were tasked with (1) identifying a K–12 mathematical problem designed to elicit an erroneous response from the LLM; (2)

formulating prompts intended to guide the model toward recognizing and resolving the error; (3) optionally posing a related problem to examine whether the LLM could apply the corrected concept; and (4) submitting the complete interaction transcript.

Data collection and analysis

The data consisted of 20 PT assignments. Guided by Systemic Functional Linguistics, the clause was selected as the unit of analysis, as it constitutes the primary locus of meaning-making in SFL, where interpersonal, ideational, and textual meanings are simultaneously realized. Each assignment was parsed into teacher turns in interaction with the LLM, and each teacher turn was further segmented into clauses, yielding a final corpus of 144 clauses.

Data analysis proceeded in two stages. First, we conducted a micro-level analysis in which each clause was coded using an operational MRT-SFL scheme. We then conducted a macro-level analysis to identify recurring patterns and trends across the dataset. The MRT-SFL coding scheme was developed iteratively, informed by SFL theory and refined in the context of teacher–LLM interaction.

Drawing on the interpersonal metafunction, we coded clauses for Mood and Modality to examine how PTs positioned themselves in interaction and opened or constrained dialogic space. Clauses were coded for Mood type as imperative (commands or direct requests for action; coded 1), declarative (statements, such as providing feedback; coded 2), closed interrogative (questions seeking specific information or a single determinate answer; coded 3), or open interrogative (questions inviting reasoning or elaboration; coded 4). Modality was coded to reflect teachers' epistemic stance toward mathematical claims, distinguishing between high epistemic Modality (coded 1), which expressed certainty or authority and positioned the teacher as the arbiter of correctness, and low/median epistemic Modality (coded 2), which introduced tentativeness or invited reconsideration and further reasoning. For the distributional analyses in the Findings, open interrogatives with low/median epistemic Modality were coded as "open", and all others as "closed."

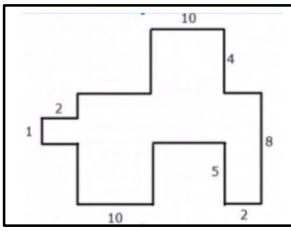
Drawing on the textual metafunction, we coded Theme as an indicator of PTs' uptake. Specifically, we examined whether the topical Theme of a teacher's clause following an LLM contribution took up the LLM's preceding mathematical reasoning. Uptake was coded as 0 when no reference occurred, 1 when the reference was partial, and 2 when the clause directly engaged with the LLM's reasoning. For distributional analyses reported in the Findings, uptake was dichotomized as no uptake (0) versus some uptake (1–2), with additional analyses reported for strong uptake (2).

To support reliability, we conducted a calibration process. Five cases were randomly selected and independently coded by all three authors using the MRT-SFL scheme. Agreement was high across categories, supporting the coding definitions. After alignment and refinement of operational definitions, the remaining cases were randomly distributed among the authors for independent coding.

FINDINGS

We begin with a fine-grained analysis of a single illustrative case to show how mathematically responsive (and non-responsive) interaction is enacted in teacher–LLM dialogue, examining how Mood, Modality, and Theme uptake interact. We then broaden to a cross-case analysis across the full dataset. For orientation, we treat PTs' prompts as forming four interactional configurations based on combining two options for Mood (open vs. non-open) and two for Theme uptake (no uptake vs. some uptake); the case below illustrates these configurations before we summarize their distribution in Table 2. To ground the typology, we analyze one participant's task (Figure 1) and selected prompts with brief summaries of the LLM's outputs.

Figure 1: Mishi's task and first prompt



"Calculate the area of the attached shape (ignore the actual scale and refer only to the measurements written on the polygon's sides. Do not measure at all. Refer only to the shape in general as given)."

As Figure 1 shows, Mishi asked the LLM to calculate the area of a polygon. The LLM described the figure as "a step-like rectilinear polygon" that could be decomposed into rectangles, then proposed enclosing it in an outer rectangle and subtracting indentations, and provided the answer 145 square units (incorrect). Mishi responded, "How do you calculate the area of a shape composed of rectangles?". This clause was coded as an open interrogative with weak uptake, echoing the LLM's "rectangles" framing. In Modality terms, it adopted low/median epistemic Modality by treating the solution as negotiable and in need of explanation. However, it did not take up the LLM's specific strategy (enclose-and-subtract) or locate the error.

Next, the LLM provided a general explanation of how to compute area by decomposing a figure into rectangles. Mishi then reprompted, "Answer the previous question using decomposition into rectangles." The LLM retried but again produced the same incorrect result. Mishi followed with a more explicit directive: "Compute the area using the standard method of dividing into individual rectangles and adding their areas." Both prompts were coded as non-open imperatives with high epistemic Modality (positioning Mishi as the authority arbiter of the appropriate approach) and no uptake: Rather than engaging the substance of the LLM's prior reasoning, the prompts constrained the model toward a preferred procedure. Consequently, they did not surface which specific step in the LLM's work produced the error—information that could have supported a more diagnostically targeted repair. After another unsuccessful attempt in which the LLM again enclosed-and-subtracted, Mishi reacted: "I am asking you to use the method of adding simple rectangles, and you are using a subtraction-of-areas method". We coded the last clause as a declarative with high epistemic Modality and some uptake, because it explicitly named the method the LLM used.

Toward the end of the exchange, as Mishi attempted to help the LLM locate the source of its mistake, she asked again, “How do you divide a complex shape into individual rectangles?” This prompt was coded as an open interrogative with low/median Modality but with no uptake, because it did not build on the LLM’s preceding contribution and largely returned to a general procedural query that the LLM had already addressed earlier in the conversation.

Notably, toward the close of the interaction, Mishi also produced several open prompts with strong uptake (coded as “2”) that directly engaged the LLM’s specific output and targeted the source of error. For example, she asked, “Explain how you previously got 145?” and “Why did you assume there was a gap in the middle?”, explicitly referencing the earlier incorrect result and interrogating the assumption that led to it. In these turns, Mishi shifted from general procedural prompting to diagnostically focused questions that pressed for justification and pinpointed the error. To sum up, this exchange illustrates MRT moves, and lack thereof: In some cases, Mishi’s open prompts were generic and did not build on the LLM’s immediate reasoning, while toward the end, she produced strongly contingent questions that explicitly took up the LLM’s output, pressed for justification, and pinpointed the source of error.

We now shift to a cross-case analysis of the full dataset. For parsimony, we focus on Mood and Theme, as Mood and Modality consistently patterned together in the analysis, with Modality not yielding additional distinctions beyond those captured by Mood. Table 2 summarizes the distribution of PT clause types and their Theme uptake of the LLM’s immediately preceding contribution. Across the corpus, uptake was evenly distributed: 72 clauses exhibited no uptake, and 72 clauses exhibited some uptake (levels 1–2). This overall balance, however, obscures systematic variation by Mood. Although open interrogative clauses accounted for 36% of the corpus, they were more likely to enact uptake than non-open clause types (closed interrogatives, imperatives, and declaratives). Specifically, approximately 60% (31/52) of the open interrogatives took up the LLM’s prior reasoning, compared with 41 of the 91 non-open clauses (approximately 45%). Thus, while non-open clauses were more prevalent overall (64%), open interrogative clauses more frequently functioned to build contingently on the LLM’s immediately prior contribution.

Table 2: PT clause Mood by Theme uptake in response to the LLM.

	Theme: Some uptake	Theme: No uptake	Total
Mood: Open	31 (~22%)	21 (~15%)	52 (~36%)
Mood: Closed interrogative/ imperative/declarative	41 (~29%)	51 (~36%)	92 (~64%)
Total	72 (~50%)	72 (~50%)	144 (100%)

Notably, open interrogatives with uptake were concentrated among a subset of

participants. Although Table 2 shows 31 such prompts, they were produced by only nine of the 20 PTs, indicating that open questions that explicitly built on the LLM's immediately preceding contribution were unevenly distributed and clustered within fewer than half of the conversations.

To further differentiate MRT levels, we next examine the strongest form of uptake (Theme uptake = 2), which captures prompts that explicitly take up LLM reasoning. In Table 2, some uptake aggregates both partial uptake (1) and strong uptake (2). At the clause level, strong uptake accounted for 25% (36/144), indicating that this tightly contingent form of engagement was present but not dominant. Only 12 of the 20 PTs produced at least one strong-uptake prompt.

DISCUSSION

This exploratory study examines baseline PT–LLM interactions to characterize MRT as reflected in PTs' spontaneous linguistic choices when engaging with an artificial agent's mathematical errors. Following Marco and Stylianides (2025), we treated LLM errors as pedagogical leverage and as a space for PTs to practice agency-preserving responses to mathematical mistakes. Using an SFL-informed operationalization (the MRT–SFL scheme), we found substantial variation in how PTs managed interactional space and built contingently on the LLM's contributions. Prompts clustered into four types (open/non-open × uptake/no-uptake), with uptake evenly distributed overall but more likely in open interrogatives. However, openness alone did not ensure contingent engagement, as some open prompts functioned as procedural resets rather than responses to the LLM's mathematical ideas. Notably, this variation may reflect differing orientations toward the LLM itself: some PTs interacted with the model as a dialogic partner whose reasoning warranted interpretation and response, whereas others treated the LLM as a tool that functions most effectively when directed through imperative prompts. More tightly contingent forms of engagement were concentrated among a subset of participants, indicating that responsiveness was available as a conversational resource but not uniformly enacted. Rather than prescribing a singular vision of MRT, this work makes visible specific interactional moves that can be productively targeted in future design iterations. It further contributes to the emerging literature on teacher–LLM mathematical conversations (Taani & Alabidi, 2024; Schwarts & Czaczkes, submitted; Zhuang & Zhang, 2025) by demonstrating the analytic utility of MRT constructs.

A key limitation of this study is that PTs interacted with a general-purpose LLM rather than a pedagogically designed simulator approximating K–12 student reasoning, which may have largely shaped the discourse. Notwithstanding, the study provides a necessary empirical baseline to inform the development of such pedagogical tools. The task structure—eliciting, interpreting, and remediating AI-generated errors—offers a flexible platform for future research extending beyond responsiveness to areas such as problem posing and error detection. Future work might also focus on the design of pedagogical simulators that better approximate student logic and interactional cues

(e.g., Schwarts & Czaczkes, submitted), enabling studies that examine whether targeted supports can shift PTs' moves toward more contingent, mathematically responsive engagement.

References

Bishop, J. P., Hardison, H. L., & Przybyla-Kuchek, J. (2022). Responsiveness to students' mathematical thinking in middle-grades classrooms. *Journal for Research in Mathematics Education*, *53*(1), 10–40. https://doi.org/10.5951/jresematheduc-2020-0188

Buber, M. (1958). *I and Thou*. Scribner.

Ergene, Ö., & Çaylan-Ergene, B. (2025). Pre-service mathematics teachers' and engineering students' perceptions of ChatGPT in mathematics: Development, validation and implementation study. *Digital Experiences in Mathematics Education*. 1–27 [online first] https://doi.org/10.1007/s40751-025-00176-x

Jacobs, V. R., Lamb, L. L., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, *41*(2), 169–202. https://doi.org/10.5951/jresematheduc.41.2.0169

Halliday, M., & Matthiessen, C. M. (2004). *Introduction to Functional Grammar* (3rd Edition edition). Hodder Arnold.

Marco, N., & Stylianides, A. J. (2025). ChatGPT's mathematical knowledge and the nature of human–AI collaboration in mathematical inquiry. *International Journal of Mathematical Education in Science and Technology*. *56*(11), 2279–2297. https://doi.org/10.1080/0020739X.2025.2543817

Rizos, I., & Gkrekas, N. (2025). The impact of LLMs on mathematics education and research at the university. *Social Sciences & Humanities Open*, *12*, 101969. https://doi.org/10.1016/j.ssaho.2025.101969

Robertson, A. D., Scherr, R. E., & Hammer, D. (Eds.). (2016). *Responsive teaching in science and mathematics*. Taylor & Francis. https://doi.org/10.4324/9781315689302

Schwarts, G. (2025). Moving beyond moves: Redefining mathematically responsive teaching as a humanistic stance. In C. Cornejo, P. Felmer, D. M. Gómez, P. Dartnell, P. Araya, A. Peri, & V. Randolph (Eds.). *Proc. 48th Conf. of the Int. Group for the Psychology of Mathematics Education* (Vol. 2, pp. 267–274). PME.

Schwarts, G., & Czaczkes, S. (Submitted). Implicit Formative Feedback in AI-Based Simulations for Assessing Mathematically Responsive Teaching. Submitted to *the second Feedback and Assessment in Mathematics Education conference* (FAME2).

Taani, O., & Alabidi, S. (2024). ChatGPT in education: Benefits and challenges of ChatGPT for mathematics and science teaching practices. International Journal of Mathematical *Education in Science and Technology*, 1–30. https://doi.org/10.1080/0020739X.2024.2357341

Zhuang, Y., & Zhang, S. (2025). Integrating ChatGPT in Mathematics Teacher Education: AI-Based Simulation Role-Playing to Support Practice-based Teaching. *International Journal of Artificial Intelligence in Education*, 1–23.