Article

Check for updates

Volume 24 (2026), pp. 157-176
American Journal of STEM Education: Issues and Perspectives
eISSN 30.3-1190| Print ISSN: 3069-0072
https://doi.org/10.32674/nyqmx528

ChatGPT's Pedagogical Approach and the Potential for Hidden Curriculum through School Context Proxies

Laurie H. Rubel
https://orcid.org/0000-0002-8032-5846
University of Haifa, Israel

Shimrit Goidel
https://orcid.org/0009-0003-3800-0428
University of Haifa, Israel

ABSTRACT

As large language models become increasingly used by teachers for lesson planning, questions arise about their nature and quality. We explore 45 mathematics lesson plans created by ChatGPT-4, across three conditions: a baseline prompt, a "Successful School" context, and a "Struggling School" context. We focus on how the lesson plans dictate the teacher's role and use this to locate the lessons on a dialogic-to-direct instruction continuum. We examine how school context descriptors in the prompt shifted its output. The lesson plans favored direct instruction across all conditions, with noteworthy shifts in response to school descriptors. The "Successful School" prompt generated lessons with additional mathematics concepts and advanced technology integration. The "Struggling School" prompt produced lessons with more monitoring and reinforcement, and, distinctively, generated attention to socioemotional dimensions of learning. We discuss these findings in terms of the potential to replicate patterns of educational stratification.

Keywords: Generative AI, mathematics education, equity, chatbots

INTRODUCTION

Teachers are increasingly turning to Large Language Models (LLMs) to assist with instructional planning, yet the pedagogical quality and nature of LLM-generated materials are not fully understood (Kim et al., 2025; Pepin et al., 2025). We focus on school mathematics, since it hold distinctive pedagogical traditions and a well-documented role as a site of educational inequality. Our central aim is to explore the pedagogical approach communicated in LLM-generated mathematics lesson plans and where that orientation falls along a continuum from direct to dialogic instruction. Further, we ask whether this orientation shifts in response to implicit contextual cues about school settings, in ways that may reinforce patterns of educational stratification.

We report the results of a bounded experiment in which we analyzed the output of a single LLM (ChatGPT-4) in response to a series of related prompts for eighth-grade probability lesson plans. This scope is exploratory, with the goal of exploring the use of LLMs as pedagogical supports and laying groundwork for a broader program of research. We created three variations of the same base prompt: an initial default prompt containing no contextual information and two variations that added distinct school context labels; all versions of the prompt referred to on-grade level classrooms. Thus, in addition to a sense of a baseline pedagogical approach, we observed the effects of adding school context information to the prompt.

The two school context labels we used intentionally leverage common, albeit problematic, policy-inflected descriptors used in the United States: a "Successful School" and a "Struggling School." These descriptors function as coded language that reflects racial and economic segregation in American schooling; schools designated as "struggling," "high-needs," "failing," or "needing support" are far more likely to serve predominantly minoritized student populations and schools designated as "successful" are more likely to serve predominantly white or affluent students (Rothstein, 2015). We investigated the extent to which these descriptors guide the LLM to differentiate its pedagogical approach. Although these specific labels are grounded in educational discourse from the United States, the underlying phenomenon, of how LLMs might respond to contextual cues in ways that may reproduce educational stratification, is relevant everywhere that teachers use LLMs for lesson planning.

HIDDEN CURRICULUM AND LLMs

The concept of hidden curriculum originates with Apple (1971, 1980), who argued that schools and schooling reproduce the ideological and economic structures of society. Apple put forth that schools and schooling reflect the interests of dominant groups with a hidden curriculum that tacitly prepares students for their place in a hierarchical social order. Anyon (1980), through her study of elementary schools in the United States, illustrated hidden curriculum in action. The schools in working-class communities tended to emphasize rote learning, compliance, and following directions, whereas the schools serving affluent communities promoted autonomy, creativity, and independent thinking. Even though these studies are decades old and were conducted in the United States, the phenomenon of educational stratification extends around the world (e.g., Archer et al., 2007).

Educational technology, and now GenAI, have emerged as powerful mediators of learning and curriculum. Since LLMs are trained on historically situated datasets, predominantly in English, they are likely to mimic the patterns embedded in those data (Bender et al., 2021). In this sense, LLMs can harbor a hidden curriculum of their own, as Warr and Heath (2025) have cautioned. One way that this operates is through differential responses to descriptors that carry implicit social meanings. These models have been trained to reject differentiation based on explicit descriptors invoking race or gender. However, emerging studies show that language that invokes race, gender, or socioeconomic class in implicit ways cues bias in these models.

One example comes from Etgar et al. (2024), who studied financial advice generated by LLMs. They discovered that embedding implicit cues about gender in their prompts produced different kinds of outputs. Although the financial circumstances were otherwise identical, ChatGPT-4 offered more cautious, simplified advice to prompts that signaled women users using an occupation that is stereotypically gendered as feminine (e.g., nursery schoolteacher), than it did to prompts that signaled men users (e.g., construction worker). Further experimental evidence comes from Warr et al. (2025). In their study with ChatGPT-3.5, they showed that the model scored writing higher when it was associated with a student attending an "elite, private school," an indirect proxy for race and socioeconomic class, than when associated with a student from an "inner-city public school." These findings establish that LLMs reflect and ultimately reproduce stratified patterns when prompted with coded language. We used similar techniques in this study to explore whether analogous patterns emerge in the generation of mathematics lesson plans.

PEDAGOGICAL APPROACH

Pedagogical approach is Remillard and Kim's (2020) term for a curriculum's pervasive yet implicit messaging about how students are to learn: how they are to

engage with mathematics, with one another, and with the teacher and how teachers are to facilitate that learning. Remillard and Kim (2020) mapped the pedagogical approaches of several elementary textbooks, along a continuum, with direct instruction and dialogic instruction at opposite poles. In this section, we draw on Munter et al.'s (2015) characterizations of these two instructional models and elaborate them in terms of their underlying theoretical frameworks.

Both models aim to develop procedural and conceptual understanding, regard practice as integral, value mathematically rigorous tasks, and conceptualize the monitoring of student reasoning as fundamental to teaching. However, the models bear essential differences regarding their underlying theoretical foundations and their approaches to peer collaboration, mathematical progression, task design, and nature and timing of feedback. Second, we examine the equity implications inherent to each approach. Finally, we discuss the application of this framework to the evaluation of LLM produced lesson plans in mathematics and how this connects to the concept of hidden curriculum.

Direct Instruction

Direct instruction holds its theoretical foundations in behaviorism and cognitivism. This approach posits that students learn by observing clear teacher-led demonstrations, practicing problems with increasing levels of difficulty, and having any errors readily corrected (Adams & Engelmann, 1996). According to this model, teaching is thus organized as the systematic transmission of knowledge through explicit teacher modeling, followed by guided practice with feedback, and an emphasis on independent practice for reinforcement. Group work is optional and if used, is typically used as a configuration for additional practice. Mathematical authority, meaning who determines if a solution is correct or who decides which tools to use, resides primarily with the teacher and the textbook (Munter et al., 2015).

Dialogic Instruction

Dialogic instruction, instead, frames mathematics learning as participation in disciplinary discourse. Its theoretical foundations are in constructivism and sociocultural theories of learning, particularly the Vygotskian view of knowledge as socially mediated and developed through language. From a dialogic perspective, learning occurs as students persevere through solving unfamiliar problems collaboratively, without teacher specific directiveness about exactly how to solve problems. Students primarily engage in comparing solution methods, making conjectures, and justifying reasoning, to build shared understanding. Key teacher moves in this perspective include eliciting student thinking, pressing students for explanations, connecting student contributions to disciplinary ideas, and sequencing student work for public evaluation (Stein et al., 2008). Feedback is aimed at advancing student mathematical authority to be able to evaluate the correctness of their own and others' reasoning (Munter et al., 2015).

Which Model Better Supports Equity?

Both direct and dialogic instruction have been claimed as the more equitable approach, and both have been criticized for reproducing inequality in different ways. These tensions are entwined with the history of the "math wars" in the United States and with broader concerns about power and authority in mathematics education (Louie & Rubel, 2020). The Standards Movement (NCTM, 1989, 2000) advocated dialogic instruction as part of reforms to mathematics education. Proponents highlighted its democratization of access to mathematical practices like justification and collaboration (e.g., Hiebert et al., 1996; Lampert, 1990; Stein et al., 2008) and argued that these support ideals of democracy (Michaels et al., 2008). Critics, however, challenged dialogic instruction as undermining mathematical rigor and precision (e.g., Klein, 2003). This ongoing polarization positions dialogic instruction as a progressive innovation and direct instruction as a conservative safeguard (Schoenfeld, 2004).

Mapping these models onto questions of equity is far from clear-cut. On one hand, proponents of direct instruction emphasize its effectiveness, particularly for students who are having difficulties (Stockard et al., 2018). Direct instruction initiatives, including the original Direct Instruction program (Engelmann & Becker, 1978), were lauded as effective in so-called "high poverty communities" (Stockard et al., 2018) and with English Language Learning students (Stedy & Alfanta, 2024). At the same time, the alignment of direct instruction with these specific student populations cannot be separated from the segregated landscape of schools in the United States. In these contexts, direct instruction runs the risk of devolving into what Haberman (1991/2010) characterized as a "pedagogy of poverty," or a cycle of rote instruction with a paired emphases on student compliance and teacher control. Thus, although its proponents see direct instruction as a lifeline to foundational skills, critics have contended that its focus on basic skills and procedures is emblematic of a two-tiered educational system, wherein marginalized youth are taught to follow rules while privileged youth are given opportunities to develop critical reasoning and argumentation skills (Rubel, 2017).

Dialogic instruction, on the other hand, is often championed as the more equitable alternative, thought to democratize intellectual authority and encourage broad participation (e.g., Diversity in Mathematics Education, 2007). However, studies have challenged this premise by revealing how dialogic learning environments can reproduce or exacerbate inequities in unexpected ways. Participation opportunities can be unevenly distributed, with high-status students dominating discourse, as shown by Langer-Osuna (2011) and Reinholz et al. (2022). Problem-based learning around real-world contexts, a hallmark of dialogic instruction, can inadvertently disadvantage working-class students by introducing unfamiliar cultural references and expectations (Lubienski, 2000). Perhaps most importantly, our understanding of dialogic instruction's equity potential remains

limited precisely because it is implemented less often in classrooms than direct instruction (O'Connor et al., 2017; Resnick et al., 2018).

Research Questions

These ongoing debates are preserved in curriculum documents, policy reports, and academic journals and constitute the data used to train LLMs (Pepin et al., 2025). Because dialogic instruction is implemented far less frequently in actual classrooms than direct instruction, the training data likely reflects a predominantly teacher-centered baseline. Indeed, emerging studies have noted that under baseline prompting conditions, LLM-generated lesson plans in mathematics tend to be teacher-centered rather than student-centered (Gurl et al., 2025; Walkington, 2025) and organized around low-cognitive-demand tasks (Sapkota & Bondurant, 2024). For example, Cameron and Mesiti (2024) found that ChatGPT-generated lessons followed a standard structure consisting of an introduction of key procedures or concepts, demonstration of key skills with examples, and then additional problems for students to practice, with limited support for important mathematical practices like reasoning, argumentation, or justification.

Furthermore, if training data reflects a stratified educational landscape, where direct instruction is the norm for schools in marginalized communities and dialogic instruction is reserved for the affluent, the LLM may have been trained to pair specific pedagogical approaches with specific school contexts. Thus, we pose the following research questions:

1. What is the pedagogical approach of LLM-generated mathematics lesson plans, as determined by how they define the roles for the teacher?
2. How, if at all, does this pedagogical approach shift when prompts include contextual cues specifying different school settings?

METHODOLOGY

The goal of our analysis was to characterize the generated lessons and locate their pedagogical approach along the direct-to-dialogic instruction continuum. We selected ChatGPT-4 (OpenAI, 2023; GPT-4-turbo) because of its prominence in current discussions regarding GenAI in education. We used temporary chat sessions with the memory feature disabled for each prompt (in December 2024). We adopted the outline of an initial prompt from Li et al. (2023), and designed a baseline default prompt that positioned the AI as an educational professional and specified key parameters:

> You are a professional in the field of mathematics education, possessing extensive teaching experience and expertise. You have the ability to integrate educational theory with practice and create instructive and actionable lesson plans to promote effective student learning. Please write a mathematics teaching unit about probability for Grade 8 learners, on

grade level. The unit should include five lessons. Provide highly detailed lesson plans, including an explanation about how long every part should take and be handled by the teacher and the expected role of the students. The classroom has a computer and a projector.

Two additional prompts added contextual information to this baseline prompt to create the "Successful School" and "Struggling School" conditions. The "Successful School" condition repeated the default prompt with this sentence at the end: "The students attend a successful school where state math exam scores are high, and all the graduates continue to four-year colleges." The "Struggling School" condition repeated the default prompt but culminated with this sentence: "The students attend a struggling school where state math exam scores are low, and most graduates do not continue to four-year colleges." These terms, while ostensibly descriptors of institutional performance, carry social and political weight in everyday and educational discourse. "Struggling Schools" are typically associated with marginalized communities and students from minoritized groups, whereas "Successful Schools" are linked with affluent, predominantly white communities (Rothstein, 2015). We ran each of these three prompts three times, always in independent sessions, to account for the stochastic nature of LLM outputs. This iterative process yielded nine distinct unit plans and resulted in a dataset of $n = 45$ individual lesson plans.

We acknowledge that terms like "successful" and "struggling" to qualify schools are socially constructed and are known to carry deficit-oriented histories. We do not use these labels with the intention of describing actual students, teachers, or schools, but, instead, to examine how GenAI systems respond to these as cues. Our analysis is focused on the LLM's responses to these labels, and whether and how these labels shifted the organization of learning and the prescribed roles of the teacher.

Data Analysis

We first reviewed the lesson plans with content analysis. This involved an iterative, comparative reading of the units to identify emergent themes, in which we noted patterns of differentiation across three domains: (1) mathematics content, (2) technology integration, and (3) assessment. In an accompanying stage of analysis, we focused on the pedagogical actions indicated for the teacher. Our choice to focus analytically on teacher action verbs is because verbs encode pedagogical stance by signaling what the teacher is supposed to do to support every lesson segment and thus serve as indicators of how teaching is imagined unfolding. This allowed us to compare the sets of lesson plans across prompt conditions and focus on how pedagogical roles are constructed in the lesson plans.

Using each line of the lesson plans, we identified all verbs that specified what the teacher was directed to do. This process produced 368 total occurrences

of verbs specifying teacher roles (119 in the Default condition, 108 in the "Successful School" condition, and 141 in the "Struggling School" condition). We began with initial open coding on this set of verbs to allow themes to emerge directly from the data. We then iteratively refined these emergent codes into final categories, which we structured according to our main research goal of locating the pedagogy on a continuum between direct instruction and dialogic instruction.

The five categories of teacher actions are detailed in Table 1. The first two categories, Show or Explain and Discuss or Elicit, represent the primary modes and distinguish between teacher-led transmission and student-contributed discourse. The third category, Monitor, Support, or Assess, includes summative forms of assessment as well as formative assessment, including those moments where the teacher was to gauge student progress and provide targeted feedback. The final two categories, Managing and Classroom as Learning Environment, account for the logistical and socioemotional dimensions of instruction respectively. The second author independently coded a representative subset of the verbs (n = 176, approximately 48% of the set of teacher verbs). We achieved an initial inter-coder agreement of 86%. We then resolved discrepancies through collaborative discussion and applied the final coding scheme to the complete dataset.

Table 1

Categorization of Instructional Action Verbs

Category	Definition	Examples
Show or explain	Instances where the teacher was to model, display, or present mathematical ideas or define terms and connect ideas.	Assign, clarify, define, demonstrate, explain, model, solve
Discuss or elicit	Instances where the teacher was to prompt, elicit, or orchestrate contributions from students to facilitate collective meaning-making.	Pose a question, facilitate discussion, guide, brainstorm, ask
Monitor, support, or assess	Instances where the teacher was to check for understanding, monitor progress, or provide real-time scaffolding during tasks.	Monitor, administer a quiz, circulate, assist, help, check

Manage	Instances where the teacher was to direct logistics, distribute materials, or organize student groupings.	Distribute, collect, group, organize, transition
Attend to socioemotional aspects	Actions oriented toward the socioemotional climate and the encouragement of mathematical mindsets.	Celebrate, encourage, challenge

RESULTS

The Baseline Pedagogical Approach

The LLM defaulted to a recognizable direct instruction model across all three prompt conditions. The lesson structure followed a roughly consistent formula, typically opening with a warm-up activity, moving through concept introduction and guided practice, and closing with independent practice and a summary. This structure largely held across all three conditions, though with some structural variation in the "Struggling School" lessons that we elaborate below.

Table 2

Frequency of Teacher Verbs By Prompt Condition

Teaching Category	Default	"Successful School"	"Struggling School"
Show or explain	69 (58%)	51 (47%)	61 (43%)
Discuss or elicit	40 (34%)	31 (29%)	39 (28%)
Manage	4 (3%)	6 (6%)	5 (4%)
Monitor, support or assess	6 (5%)	19 (18%)	25 (18%)
Attending to socioemotional aspects	0 (0%)	1 (1%)	11 (8%)
Total	119 (100%)	108 (100%)	141(100%)

Across all three prompt conditions, Show or Explain verbs dominated over Discuss or Elicit verbs: 58% versus 34% in the Default condition, 47% versus 29% in the "Successful School" condition, and 43% versus 28% in the "Struggling

School" condition (Table 2). The wide gap between these two categories, with ratios ranging from 1.56 to 1.72, represents the emphasis in the lessons on direct instruction practices. The 16 verbs common to all three conditions, including explain, demonstrate, show, discuss, ask, and facilitate, reflect a baseline in which dialogic moves are present but are far outnumbered by direct instruction ones (Table 3). These findings corroborate prior research suggesting that LLM-generated mathematics lesson plans tend toward teacher-centered, procedural instruction under baseline prompting conditions.

Table 3
Common and Unique Verbs for Teaching Actions

Category	Verbs
Common to All Prompts	Ask, circulate, define, demonstrate, discuss, display, elicit, facilitate, monitor, pose, provide, recap, review, show, use, summarize
Unique to Default Prompt	Calculate, formalize, have them, practice, project
Unique to "Successful School" Prompt	Administer, assist, challenge, highlight, illustrate, transition, troubleshoot
Unique to "Struggling School"	Address, assess, build (a supportive atmosphere), connect, celebrate, encourage, invite, reinforce, spark, wrap-up

Shifts in Pedagogical Approach According to School Context Cues

Even though all three prompts specified on-grade level classrooms, the introduction of school context labels shifted the LLM's output in specific ways. A chi-square test confirmed a statistically significant association between prompt condition and teacher verb category [χ^2 (8, n = 368) = 28.68, $p < .001$]. For this analysis, their significance lies in the pattern of differentiation they reveal and less in the statistics. The "Successful School" and "Struggling School" conditions diverged from the Default and from each other in four important ways:

mathematics content and cognitive demand, assessment and monitoring, socioemotional support, and technology integration.

Mathematics Content and Cognitive Demand. One divergence concerned mathematical content. Two of the three "Successful School" units included advanced topics that were absent from all other conditions: Addition Rules for non-mutually exclusive events, formal set notation, and large-sample digital simulations. Furthermore, one "Successful School" unit included a task requiring students to critically evaluate a game for fairness, a task demanding critical mathematical literacy. The "Successful School" condition was also the only one to include student presentations of mathematical findings. None of these topics or competencies appeared in the default or "Struggling School" lessons, even though all three prompts indicated that students were on grade level. This pattern suggests that the "Successful School" descriptor activated assumptions about intellectual capacity or instructional expectations.

Assessment and Monitoring. Both context conditions directed the LLM toward sharp increases in assessment: Monitor, Support, or Assess verbs rose from 5% in the Default condition to 18% in both context conditions. However, the nature of assessment differed meaningfully between the two. In the "Successful School" lessons, formative assessment was woven into instruction as teachers were directed to circulate, monitor experiments, and provide individualized support. One unit included a summative quiz at the end of the fifth lesson. The "Struggling School" lessons, by contrast, featured a denser and more structured approach to assessment: repeated "exit tickets" prompted student reflections, gamified reviews of material through Kahoot or Jeopardy-style activities, and summative assessments at the conclusion of all three generated units. The frequency and variety of assessment in the "Struggling School" lessons suggests an implicit assumption that these students need more frequent checking and verification of learning.

Socioemotional Support. Perhaps the most unexpected divergence was in the dimension of attending to socioemotional aspects of learning. The "Struggling School" condition generated 11 teacher verbs (8% of its total) directed at the socioemotional climate of the classroom, including: build a supportive atmosphere, spark curiosity, celebrate learning progress, and provide encouragement. The Default condition produced zero such verbs, and the "Successful School" condition produced only one: challenge students. This near-exclusive association of socioemotional support with the "Struggling School" condition is concerning. That is, attending to students' emotional well-being is valuable. Yet its concentration in the "Struggling School" lessons, paired with the absence of advanced mathematics content in those lessons, suggests a decoupling of care from cognitive demand. The LLM appeared to treat emotional

support and mathematical rigor as if they were demanded by different kinds of classrooms.

Technology Integration. The integration of educational technology differed across conditions. In both the Default and "Struggling School" lessons, technology served a presentational role primarily. Teachers were directed to use projectors to display definitions or show examples, or to deploy presentation platforms for review games. The "Struggling School" lessons specifically emphasized interactive slides and short video clips as scaffolds for engagement. The "Successful School" lessons, by contrast, directed teachers toward technology as an investigative tool in mathematics, with suggestions for tools such as GeoGebra and PhET interactive simulations to enable students to analyze results and model real-world events. Notably, none of the lesson plans across any condition suggested integrating artificial intelligence technologies into eighth-grade probability instruction.

DISCUSSION

This study examined the pedagogical approaches in ChatGPT-4-generated mathematics lesson plans and explored the effects of introducing contextual cues about school settings. We organize our discussion around our two research questions before turning to the broader implications of this study's findings.

Our first research question concerned the pedagogical approach to LLM-generated mathematics lesson plans. The generated lessons emphasized teacher-centered, transmission-oriented direct instruction practices, structured around explicit modeling followed by guided and independent practice. This corroborates findings from Cameron and Mesiti (2024) and Sapkota and Bondurant (2024). The consistency of this pattern suggests that the LLM has learned a particular script for what a mathematics lesson is, which reflects the predominance of direct instruction in the data on which it has been trained. As we noted in our theoretical framework, dialogic instruction is implemented far less frequently than direct instruction in classrooms (O'Connor et al., 2017; Resnick et al., 2018), meaning that the training data available to LLMs likely reflects this imbalance. The LLM's default pedagogical approach may thus reflect the statistical weight of direct instruction in its training corpus, though the opacity of commercial LLM development makes it impossible to fully distinguish training data reflection from deliberate design choices. What makes this pattern consequential is that it encodes a pedagogical approach without signaling that a contested choice has been made.

Our second research question explored whether and how pedagogical approaches shift in response to school context descriptors. We found that the descriptors "Successful School" and "Struggling School" functioned as proxies for demographic categories that shifted the LLM's pedagogical approach in meaningful ways. This finding extends Etgar et al.'s (2024) and Warr et al.'s (2025)

findings that LLMs adjust their outputs in response to implicit demographic cues. The school context labels we used, while ostensibly descriptors of institutional performance, appear to have activated associations in the LLM's training data about the kinds of students who attend such schools, and about what those students need. Teachers who include contextual information in their prompts about their schools, such as "Title I," "high-needs," "low-SES," “language learners,” “new immigrants,” or otherwise, may unknowingly trigger these or other associations, receiving lesson plans shaped by embedded assumptions.

The pattern of differentiation revealed by our findings maps onto Anyon's (1980) illustration of hidden curriculum. Just as Anyon documented how schools in working-class communities emphasized rote compliance and following directions while schools in affluent communities promoted autonomy and independent thinking, our findings show that the LLM generated lessons emphasizing monitoring, reinforcement for the “Struggling School,” while generating lessons with advanced mathematical content, investigative technology, and student presentations for the “Successful School.” This echoes what Haberman (1991/2010) characterized as a pedagogy of poverty, through which students from marginalized groups are primarily managed and monitored, but offered fewer intellectual opportunities. That this pattern emerged from a prompt difference of a single sentence is noteworthy. It suggests that the historical stratification of American schooling is embedded in LLM training data, which can be retrieved and reproduced whenever contextual cues are present, as recently argued by Warr and Heath (2025). Future research should build on this study’s preliminary findings to explore GenAI lesson planning across lesson topics, grade levels, and AI models, investigating this question at scale. Across a range of national contexts, future research could investigate how locally circulating policy labels, the equivalents of "Struggling" and "Successful," may trigger stratified pedagogical outputs in LLM-generated lessons. The present study provides the conceptual framing and methodological approach for that broader research program.

The near-exclusive presence of socioemotional support directives in the "Struggling School" lessons is noteworthy. In these lessons, the plans directed teachers to build supportive atmospheres, celebrate small wins, and spark curiosity, while simultaneously offering less advanced mathematical content and more monitoring. In the "Successful School" lessons, by contrast, the only socioemotional verb that appeared was "challenge students," whereby care was expressed as intellectual demand rather than emotional scaffolding. This pattern suggests a false tension between mathematical rigor and socioemotional care, with the LLM treating them as belonging to different kinds of classrooms and, by extension, different kinds of students.

This distinction is problematic in two ways. First, the LLM attended to socioemotional dimensions of learning only in remedial contexts, as if emotional safety were a prerequisite to engagement and achievement rather than something that can be cultivated through it. This framing is at odds with a substantial body of

research that has shown that intellectual challenge is itself a form of care, and that high expectations, when combined with strong relational support, are among the most powerful drivers of student engagement and learning (e.g., Ladson-Billings, 1995). By reserving academic challenges for students presumed to be already succeeding, the LLM encodes a logic of lower expectations for certain students. Second, this finding surfaces a question about instruction for students thought to be successful. The near absence of socioemotional language in the "Successful School" plans implies they are emotionally self-sufficient, that belonging and motivation can be taken for granted. However, research on high-achieving students suggests otherwise: the experience of being seen, valued, and supported matters across the achievement spectrum (Bartell, 2011).

Limitations

This study and its findings are preliminary, a primary limitation being its relatively small dataset (45 lessons), only one mathematics topic, a single LLM, and around one base prompt. Future research could extend these findings across topics, grade levels, and models. A second limitation concerns the semantic confounding of the two contextual descriptors ("Successful School" and "Struggling School"). This difference in affective tone might influence generative text outputs, independent of any associations with demographic groups (Bardol, 2025). Thus, our findings may reflect effects of sentiment polarity and not necessarily the educational contexts that we intended to activate with these descriptors. A third limitation is that we treated ChatGPT's responses as static, one-off outputs, whereas in practice, teachers could refine outputs through iterative prompting. Thus, our findings capture the model's initial pedagogical approach rather than the potentially more nuanced plans that could emerge through a teacher's sustained interaction with it.

Conclusions

These findings hold implications for teacher education and for teachers who use LLMs for lesson planning. Most current use of LLMs remains limited to zero-shot interactions, single prompts without iterative refinement (Dilling & Herrmann, 2024), meaning that the assumptions embedded in the LLM's initial response are likely to go unexamined. A teacher who describes their school context in a prompt, intending to personalize the lesson, may receive a plan shaped by the LLM's encoded assumptions about what students in that context need. Supporting teachers in embedding explicit pedagogical and equity-oriented guidance in their prompts is therefore important, so that instructional decisions do not silently default to the historical patterns the LLM has absorbed. However, prompt literacy in teacher education is an incomplete solution, because the assumptions this study identified are not surface-level features that a better-worded prompt would override. They reflect patterns that may persist even when teachers attempt to redirect them. For this reason, teacher education must also support teachers in

developing critical awareness to evaluate AI-generated content on its own terms, including noticing when a lesson reflects a particular pedagogical approach and recognizing that an ostensibly neutral planning tool can carry a hidden curriculum of its own.

Acknowledgements

The authors used generative AI tools (OpenAI; Grammarly) to assist with language editing and formatting during the manuscript preparation process. All ideas, interpretations, and findings are the original work of the authors. Portions of this research were accepted for presentation at the Psychology of Mathematics Education 49 Conference.

REFERENCES

Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR*. Educational Achievement Systems. https://www.thalesacademy.org/assets/docs/DI-25_Yrs_Beyond.pdf

Anyon, J. (1980). Social class and the hidden curriculum of work. *Journal of Education*, *162*(1), 67–92. https://doi.org/10.1177/002205748016200106

Apple, M. W. (1971). The hidden curriculum and the nature of conflict. *Interchange, 2*(4), 27–40. https://doi.org/10.1007/BF02287080

Apple, M. W. (1980). The other side of the hidden curriculum: Correspondence theories and the labor process. *Interchange, 11*(3), 5–22. https://doi.org/10.1007/bf01190034

Archer, L., Hollingworth, S., & Halsall, K. (2007). 'University's not for me—I'm a Nike person': Urban, working-class young people's negotiations of 'style', identity and educational engagement. *Sociology*, *41*(2), 219–237. https://doi.org/10.1177/0038038507074731

Bardol, F. (2025). ChatGPT reads your tone and responds accordingly—until it does not—emotional framing induces bias in LLM outputs. *arXiv*. https://arxiv.org/abs/2507.21083v1

Bartell, T. (2011). Caring, race, culture, and power: A research synthesis toward supporting mathematics teachers in caring with awareness. *Journal of Urban Mathematics Education*, *4*(1), 50–74. https://jume-ojs-tamu.tdl.org/jume/article/view/128/84

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Cameron, S., & Mesiti, C. (2024). What kind of mathematics teacher is ChatGPT? Identifying the pedagogical practices referenced by generative AI tools when preparing lesson plans. In J. Višňovská, E. Ross, & S.

Getenet (Eds.), *Surfing the waves of mathematics education: Proceedings of the 46th annual conference of the Mathematics Education Research Group of Australasia* (pp. 135–142). MERGA. https://files.eric.ed.gov/fulltext/ED661134.pdf

Dilling, F., & Herrmann, M. (2024). Using large language models to support pre-service teachers' mathematical reasoning—an exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1460337

Diversity in Mathematics Education Center for Learning and Teaching. (2007). Culture, race, power and mathematics education. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 405–433). Information Age Publishing.

Engelmann, S., & Becker, W. C. (1978). Systems for basic instruction: Theory and applications. In A. C. Catania & T. A. Brigham (Eds.), *Handbook of applied behavior analysis* (pp. 325–377). Irvington Publishers.

Etgar, R., Guttmann, R., Taub-Tabib, H., & Shmueli, E. (2024). *Implicit bias in LLMs: Bias in financial advice based on implied gender* (SSRN Scholarly Paper No. 4930361). SSRN. https://doi.org/10.2139/ssrn.4930361

Gurl, T., Markinson, M., & Artzt, A. F. (2025). Using ChatGPT as a lesson planning assistant in secondary mathematics teacher education. *Digital Experiences in Mathematics Education, 11,* 114–139. https://doi.org/10.1007/s40751-024-00162-9

Haberman, M. (1991/2010). The pedagogy of poverty versus good teaching. *Phi Delta Kappan*, *92*(2), 81–87. https://doi.org/10.1177/003172171009200223

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K., Wearne, D., Murray, H., Olivier, A., & Human, P. (1996). Problem solving as a basis for reform in curriculum and instruction: The case of mathematics. *Educational Researcher, 25*(4), 12–21. https://doi.org/10.3102/0013189X025004012

Kim, D., Jung, J., & Sheumaker, M. F. (2025, March). Teacher Education Students' Perceptions of ChatGPT for Lesson Planning: Benefits, Challenges, and Use. In *Society for Information Technology & Teacher Education International Conference* (pp. 756-761). Association for the Advancement of Computing in Education (AACE).

Klein, D. (2003). A brief history of American K-12 mathematics education in the 20th century. http://www.csun.edu/~vcmth00m/AHistory.html

Ladson-Billings, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice, 34*(3), 159–165. https://www.jstor.org/stable/1476635

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal, 27*(1), 29–63. https://doi.org/10.2307/1163068

Langer-Osuna, J. M. (2011). How Brianna became bossy and Kofi came out smart: Understanding the trajectories of identity and engagement for two group leaders in a mathematics classroom. *Canadian Journal of Science, Mathematics and Technology Education, 11*(3), 207–225. https://doi.org/10.1080/14926156.2011.595881

Li, Y., Liu, J., & Yang, S. (2023). Is ChatGPT a good middle school teacher? An exploration of its role in instructional design. *Proceedings of the 3rd International Conference on New Media Development and Modernized Education*. https://doi.org/10.4108/eai.13-10-2023.2341343

Louie, N. & Rubel, L. (2020). *Opposition to change in mathematics education in the United States.* [unpublished manuscript]. School of Education, University of Wisconsin-Madison.

Lubienski, S. T. (2000). Problem solving as a means toward mathematics for all: An exploratory look through a class lens. *Journal for Research in Mathematics Education*, *31*(4), 454-482. https://doi.org/10.2307/749653

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*, 283–297. https://doi.org/10.1007/s11217-007-9071-1

Munter, C., Stein, M. K., & Smith, M. S. (2015). Dialogic and direct instruction: Two distinct models of mathematics instruction and the debate(s) surrounding them. *Teachers College Record*, *117*(11), 1–32. https://doi.org/10.1177/016146811511701102

National Center for Education Statistics. (2023). *Concentration of public-school students eligible for free or reduced-price lunch*. Institute of Education Sciences. https://nces.ed.gov/programs/coe/indicator/clb

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. National Council of Teachers of Mathematics.

O'Connor, C., Michaels, S., Chapin, S., & Harbaugh, A.G. (2017). The silent and the vocal: Participation and learning in whole-class discussion. *Learning and Instruction, 48*, 5–13. https://doi.org/10.1016/j.learninstruc.2016.11.003

OpenAI. (2023). *GPT-4-turbo* [Large language model]. https://platform.openai.com/docs/models

Pepin, B., Buchholtz, N. & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education, 11*, 9–41. https://doi.org/10.1007/s40751-025-00172-1

Reinholz, D., Johnson, E., Andrews-Larson, C., Stone-Johnstone, A., Smith, J., Mullins, B., Fortune, N., Keene, K., & Shah, N. (2022). When active learning is inequitable: Women's participation predicts gender inequities in

mathematical performance. *Journal for Research in Mathematics Education*, *53*(3), 204–226. https://doi.org/10.5951/jresematheduc-2020-0143

Remillard, J. T. & Kim, O. K.,(2020). *Elementary mathematics curriculum materials: Designs for student learning and teacher enactment*. Springer. https://doi.org/10.1007/978-3-030-38588-0

Resnick, L. B., Asterhan, C. S. C., Clarke, S. N., & Schantz, F. (2018). Next generation research in dialogic learning. In G. E. Hall, L.F. Quinn, & D.M. Gollnick. (Eds.), *Wiley handbook on teaching and learning* (pp. 323–338). Wiley-Blackwell. https://www.wiley.com/en-us/The+Wiley+Handbook+of+Teaching+and+Learning-p-9781118955871

Rothstein, R. (2015). The racial achievement gap, segregated schools, and segregated neighborhoods: A constitutional insult. *Race and Social Problems, 7*(1), 21–30. https://doi.org/10.1007/s12552-014-9134-1

Rubel, L.H. (2017). Equity-directed instructional practices: Beyond the dominant perspective. *Journal of Urban Mathematics Education, 10*(2), 66–105. https://doi.org/10.21423/jume-v10i2a324

Sapkota, B. & Bondurant, L. (2024). Assessing concepts, procedures, and cognitive demand of ChatGPT-generated mathematical tasks. *International Journal of Technology in Education* (IJTE), *7*(2), 218–238. https://doi.org/10.46328/ijte.677

Schoenfeld, A. H. (2004). The math wars. *Educational Policy, 18*(1), 253–286. https://doi.org/10.1177/0895904803260042

Stedy, Y., & Alfanta, D. (2024). The influence of the Direct Instruction learning model on student learning outcomes in elementary schools. *Journal of Education Innovation and Curriculum Development*, *2*(3), 99–111. https://journals.iarn.or.id/index.php/educur/article/view/446

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, 10*(4), 313–340. https://doi.org/10.1080/10986060802229675

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research, 88*(4), 479–507. https://doi.org/10.3102/0034654317751919

Trust, T., Maloy, R., Xu, C., & Pelletier, K. (2025). Civic education in the age of AI: Should we trust AI-generated lesson plans? *Contemporary Issues in Technology and Teacher Education, 25*(3). https://citejournal.org/volume-25/issue-3-25/social-studies/civic-education-in-the-age-of-ai-should-we-trust-ai-generated-lesson-plans

Walkington, C. (2025). The implications of generative artificial intelligence for mathematics education. *School Science and Mathematics*, 1–10. https://doi.org/10.1111/ssm.18356

Warr, M., & Heath, M. K. (2025). Uncovering the hidden curriculum in generative AI: A reflective technology audit for teacher educators. *Journal of Teacher Education*, *76*(3), 245–261. https://doi.org/10.1177/00224871251325073

Warr, M., Oster, N. J., & Isaac, R. (2025). Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education*, *57*(6), 1324–1349. https://doi.org/10.1080/15391523.2024.2395295

Bios

LAURIE H. RUBEL, PhD, is an Associate Professor in the Faculty of Education at the University of Haifa. Her research interests include mathematics education, teacher education, and educational equity. Email: LRubel@edu.haifa.ac.il

SHIMRIT GOIDEL is a high-school mathematics teacher and graduate student at the University of Haifa. Email: shimrit.goidel@gmail.com