Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 24-1028

Scaffolding Middle-School Mathematics Curricula With Large Language Models

Rizwaan Malik
Stanford University

Dorna Abdi
Stanford University

Rose Wang
Stanford University

Dorottya Demszky
Stanford University

Despite well-designed curriculum materials, teachers often face challenges implementing them due to diverse classroom needs. This paper investigates whether large language models (LLMs) can support middle school math teachers by helping create highquality curriculum scaffolds, which we define as the adaptations and supplements teachers employ to ensure all students can access and engage with the curriculum. Through cognitive task analysis with expert teachers, we identify a three-stage process for curriculum scaffolding: observation, strategy formulation, and implementation. We incorporate these insights into three LLM approaches to create warmup tasks that activate students' background knowledge. The best-performing approach provides the model with the original curriculum materials and an expert-informed prompt; this approach generates warmups that are rated significantly higher than those created by expert teachers in terms of alignment to learning objectives, accessibility to students working below grade level, and teacher preference. This research demonstrates the potential of LLMs to support teachers in creating effective scaffolds and provides a methodology for developing artificial intelligence-driven educational tools.

VERSION: April 2025

Suggested citation: Malik, Rizwaan, Dorna Abdi, Rose Wang, and Dorottya Demszky. (2025). Scaffolding Middle-School Mathematics Curricula With Large Language Models. (EdWorkingPaper: 24 -1028). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/b47y-mh41

Scaffolding Middle School Mathematics Curricula With Large Language Models

Rizwaan Malik*

Dorna Abdi†

Rose Wang‡

Dorottya Demszky§

Abstract: Despite well-designed curriculum materials, teachers often face challenges implementing them due to diverse classroom needs. This paper investigates whether large language models (LLMs) can support middle school math teachers by helping create high-quality curriculum scaffolds, which we define as the adaptations and supplements teachers employ to ensure all students can access and engage with the curriculum. Through cognitive task analysis with expert teachers, we identify a three-stage process for curriculum scaffolding: observation, strategy formulation, and implementation. We incorporate these insights into three LLM approaches to create warmup tasks that activate students' background knowledge. The best-performing approach provides the model with the original curriculum materials and an expert-informed prompt; this approach generates warmups that are rated significantly higher than those created by expert teachers in terms of alignment to learning objectives, accessibility to students working below grade level, and teacher preference. This research demonstrates the potential of LLMs to support teachers in creating effective scaffolds and provides a methodology for developing artificial intelligence-driven educational tools.

Keywords: curriculum scaffolding; middle school mathematics; human–computer interaction; large language models; cognitive task analysis; large language model evaluation

Practitioner's Notes

What is already known about this topic

*Rizwaan Malik (rizmalik@stanford.edu; corresponding author) is a Graduate Researcher in Education Data Science at Stanford University, Stanford, CA, United States.

†Dorna Abdi (dabdi@stanford.edu) is a Graduate Researcher in Education Data Science at Stanford University, Stanford, CA, United States.

‡Rose Wang (rewang@stanford.edu) is a Ph.D. student in Computer Science at Stanford University, Stanford, CA, United States.

§Dorottya (Dora) Demszky (ddemszky@stanford.edu; corresponding author) is an Assistant Professor in Education Data Science at Stanford University, Stanford, CA, United States.

1. Scaffolding is essential for enabling students to access and engage with curriculum materials.
2. Large language models (LLMs) have shown promise in generating educational content and supporting teachers.
3. Teachers frequently need to adapt and supplement standardized curricula to meet the diverse needs of their students.

What this paper adds

1. Identifies a three-stage curriculum scaffolding process (observation, strategy formulation, implementation) used by expert teachers.
2. Demonstrates that providing LLMs with additional context from the curriculum, such as the original warmup task, helps to ground the model and improve the quality of the generated warmup tasks.
3. Demonstrates that, when prompted well, LLMs can generate warmup tasks that are of similar or better quality compared to those created by expert teachers in terms of alignment to learning objectives, accessibility, and teacher preference.

Implications for practice and/or policy

1. Provides practical suggestions for prompting LLMs to generate high-quality warmup tasks for middle school math teachers, such as incorporating additional curriculum context and expert-informed prompts.
2. Demonstrates how cognitive task analysis with expert teachers can be used to develop LLM-based tools for educators that align with their practices and preferences.
3. Indicates that additional research is needed to explore the potential for LLMs to support other types of curriculum adaptations, evaluate their effectiveness in classroom settings, and investigate how they can be effectively tailored to the specific needs and characteristics of individual students.

1 Introduction

Scaffolding is an instructional process that provides support tailored to students that enables them to "solve problems, complete tasks, or achieve educational goals that would be challenging without such support" (Wood et al., 1976). Research consistently shows that scaffolded instruction benefits all students, whether they have documented learning needs or not, leading to improved academic and nonacademic outcomes (CAST, 2024; Goddard et al., 2015). The need for scaffolding has been magnified in the post-pandemic era, as the national incidence of students with Individualized Education Plans (IEPs) and 504 plans has increased from 13% to 15%, with some districts reporting rates as high as 35% (Irwin et al., 2023). Furthermore, an alarming 49% of students are now testing below grade level in at least one subject (Irwin et al., 2023).

At the same time, in recent years the use of standardized curricula in math education has proliferated to support teachers in providing high-quality instruction (Doan et al., 2022; Jackson & Makarin, 2018; Remillard & Heck, 2014). However, the post-pandemic era has highlighted the limitations of these curricula in addressing the diverse needs of learners (Nerlino, 2022). With more students working below grade level or requiring special support, standardized curricula often lack the necessary scaffolding for accessibility and engagement. Curriculum scaffolding, a core aspect of instructional scaffolding, involves teachers adapting and supplementing the official curriculum to bridge this gap, particularly for struggling learners (Remillard, 1999; Remillard & Heck, 2014; Squire et al., 2003). This may include adding warmup tasks, formative assessments, or culturally relevant adaptations (Meidl & Meidl, 2011; Walkington, 2013).

Despite the documented effectiveness of and need for scaffolding, a majority of educators encounter significant challenges in implementing it, and there's little work on how teachers approach the scaffolding process (Sherin & Drake, 2009). One of the primary challenges is the time-intensive nature of creating effective scaffolds, which requires teachers to carefully analyze the curriculum, identify potential barriers for learners, and design appropriate adaptations and supplements. This process can be particularly challenging for novice teachers, who may lack the experience and pedagogical content knowledge to efficiently navigate the complexities of scaffolding (Lokey-Vega, 2015). Additionally, many teachers receive limited training on how to effectively scaffold instruction, highlighting the need for more targeted teacher education in this area (Galiatsos et al., 2019; Yan & Goh, 2023).

This research investigates the potential for large language models (LLMs) to support curriculum scaffolding in middle school mathematics. Effectively exploring this potential first requires a more granular understanding of how teachers approach the scaffolding process. Thus, we start with a qualitative research phase, conducting cognitive task analysis (CTA) to gain insight into how expert teachers scaffold their curricula. Then, building on our findings and existing frameworks (Remillard & Heck, 2014), we propose a three-stage process of curriculum scaffolding that comprises observation, strategy formulation, and implementation. We apply this framework to a case study using LLMs to generate warmup tasks—a specific type of instructional scaffold that activates and refreshes background knowledge. To assess LLMs' ability to produce warmup tasks of a similar quality as those made by experts, we develop a novel dataset of expert-created warmup tasks and prepare various LLM-based approaches informed by insights from our CTA. We then evaluate these LLM-based approaches against the expert-created tasks using considerations identified in the CTA as evaluation criteria.

Our research explored the following interdependent questions:

1. How do expert teachers scaffold their curricula, and which aspects of their process might artificial intelligence (AI) facilitate?

2. How do instructional scaffolds created by LLMs compare in quality to those created by expert teachers?

2 Related Work

2.1 Conceptions of Curriculum Use

A large area of educational research centers on curriculum use: How do teachers interact with and draw upon curriculum resources designed to guide instruction (Remillard, 2005; Stein & Smith, 1998)? Recent frameworks have sought to provide a more nuanced understanding of the curriculum implementation process by distinguishing between the official and operational curriculum (Remillard & Heck, 2014). The *official curriculum*, also known as the *formal curriculum* (Gehrke et al., 1992), comprises curricular aims and objectives, content of consequential assessments, and the designated curriculum (Remillard & Heck, 2014). In contrast, the *operational curriculum* encompasses what actually happens in the classroom, including the teacher-intended curriculum, the enacted curriculum, and student outcomes (Gehrke et al., 1992; Remillard & Heck, 2014).

The transition from the official to the operational curriculum involves teachers interpreting and making decisions when designing instruction, drawing upon the designated curriculum and other resources. This process, termed *documentational genesis*, results in the generation of new documents tailored to specific students at a particular moment (Gueudet & Trouche, 2009). Sherin & Drake (2009) proposed a three-step process of reading, evaluating, and adapting, while Brown (2002) described three modes of engagement: offloading, adapting, and improvising. This process of curriculum enactment is highly dependent on teacher characteristics, such as subject matter knowledge, pedagogical content knowledge, beliefs, and goals, as well as the local context (Brown, 2002; Gonzalez Thompson, 1984; Remillard, 2005; Sherin & Drake, 2009).

As noted by Sherin & Drake (2009), there have been relatively few empirically based generalizations focused on the transition from the official to the operational curriculum, and even fewer from the designated to the teacher-intended. This is partly due to the challenge of accessing the teacher-intended curriculum, given that it typically exists within individual teacher minds, files, and presentations, and it is not centrally collected or stored (Remillard & Heck, 2014). Our study aims to help fill this gap by examining the strategies and considerations expert teachers employ in scaffolding curriculum for diverse learners and exploring the potential for AI to support this process. By focusing on the transition from the designated curriculum to the teacher-intended curriculum and the specific adaptations and supplements teachers make to ensure accessibility and engagement for all students, we seek to provide a more nuanced understanding of curriculum enactment in contexts of learner variability.

2.2 Scaffolding in Math Education

The effectiveness of scaffolding in math education is closely tied to the transition from the official to the operational curriculum. The concept of scaffolding refers to the temporary but essential support provided to learners to help them achieve tasks that would otherwise be beyond their reach (Wood et al., 1976). Grounded in Vygotsky's sociocultural theory and the concept of the zone of proximal development (ZPD), scaffolding emphasizes the importance of providing learners with the necessary guidance and support to move from assisted to independent performance (Vygotsky & Cole, 1978). In the context of math education, scaffolding has been shown to be a highly effective instructional approach, with meta-analyses reporting significant effect sizes on student outcomes (Hattie, 2008; Zuo et al., 2023). The benefits of scaffolding extend beyond academic achievement, encompassing increased task effort, cognitive development, metacognitive awareness, independence, sensemaking, and self-confidence (Zuo et al., 2023).

Well-designed, standards-aligned instructional materials often remain inaccessible to students if the tasks fall outside their ZPD. Consequently, teachers often devote significant time and effort to adapting and supplementing their core curriculum materials to better suit their students' learning needs (Philipp & Kunter, 2013). This process of curriculum scaffolding, which we define as a subset of the teacher-intended curriculum, encompasses the various techniques and strategies teachers employ to ensure that all students can access and engage with the official curriculum. These techniques include providing hints; modeling; asking probing questions; employing the gradual release of responsibility model; utilizing guided practice, visual aids, manipulatives, graphic organizers, and supportive resources; and offering alternative representations (Drew, 2022). By strategically using these techniques, teachers can tailor their support to the diverse needs of their students, helping each learner navigate the path toward understanding and independence.

2.3 Automated Creation of Educational Materials

The rapid advancement of LLMs, such as GPT-4o, has opened up new avenues for the automated creation of educational materials. These models, trained on vast amounts of text data, have demonstrated the ability to generate human-like text in response to given inputs, offering potential applications in various educational contexts. The proliferation of AI tools and chatbots has led to increased usage among educators, with nearly 50% now using ChatGPT at least once a week (Impact Research, 2024).

Recent studies have explored the use of LLMs in generating a wide range of educational materials. LLMs have been employed to author learning objectives, ensuring alignment with course content and desired outcomes (Sridhar et al., 2023); they have also been used to write worked examples and explanations, providing step-by-step guidance for learners (Jury et al., 2024; Prihar et al., 2023). In the domain of question generation, LLMs have demonstrated the ability to create question–answer pairs, multiple-choice questions, and open-ended questions across various subjects (Bulathwela et al., 2023; Doughty et al., 2024; Elkins et al., 2023; Rodriguez-Torrealba et al., 2022; Shimmei et al., 2023; Z. Wang et al., 2022). In the context of math education, LLMs have been utilized to generate problems at varying levels of difficulty and to adapt existing problems for improved student understanding (Jiao et al., 2023; Norberg et al., 2023). LLMs have also been employed to generate content for entire courses, including syllabi, lectures, and assessments (Diwan et al., 2023; Leiker et al., 2023). This line of work suggests that LLMs could serve as powerful tools for creating instructional resources, potentially saving teachers time and effort in developing materials tailored to their students' needs.

However, the use of LLMs in educational contexts is not without challenges. Researchers have noted that these models can sometimes generate unreliable solutions to math problems (Frieder et al., 2024) and may "hallucinate" information, producing content that appears plausible but is not actually accurate (Ji et al., 2023a). There is also evidence that LLMs alone do not behave like expert instructors, such as when providing pedagogical feedback to teachers (R. Wang & Demszky, 2023), remediating mathematical mistakes in tutoring (R. E. Wang et al., 2024), or providing pedagogical explanations (Jury et al., 2024; Prihar et al., 2023). These limitations highlight the need for careful consideration and human oversight when employing LLMs in the creation of educational materials.

One promising way to ensure both safety and pedagogically sound outputs is to create AI-in-the-loop systems where the teacher is in control of what is being sent to students (Ninaus & Sailer, 2022). For example, in a study exploring the use of LLMs for remediating student math mistakes — a type of scaffolding – R. E. Wang et al. (2024) demonstrated that LLM responses to students improved significantly when the teacher helped identify the cause of error-and-response strategy to use.

2.4 Modeling Expert Decision Making

Understanding the decisions instructional experts make while performing specific tasks is critical to designing technological systems that seek to model such expertise and enhance teachers' work. CTA is a widely used, structured, qualitative research method for eliciting and formalizing the knowledge, thought processes, mental strategies, and goal structures that underlie expert performance (Clark et al., 2008). The approach builds upon and extends traditional ethnographic research methods by incorporating not only observations but also verbal statements from experts as a primary source of information. Through the use of interview techniques, verbal reports, and analysis of team communication, researchers can elicit valuable insight into the cognitive processes that guide expert performance.

In recent years, researchers have increasingly applied CTA in educational settings to gain insight into the complex decision-making processes of expert teachers. For example, Lokey-Vega (2015) used CTA to detail a nine-step process that experts follow when designing and implementing technology-rich lessons, revealing that novice teachers were less familiar with distinct parts of this process. Similarly, R. E. Wang et al. (2024) employed CTA to develop a framework for expert teacher decision making in the context of responding to student misconceptions, formalizing a process in which the expert identifies the student's error, determines a remediation strategy, and articulates their instructional intention before generating a response.

In the context of curriculum scaffolding, CTA offers a powerful tool for identifying the knowledge and strategies that expert teachers use to tailor curriculum materials and create effective scaffolds for diverse learners. By formalizing these often tacit processes, CTA can help make expert teachers' decision making more explicit and accessible, potentially informing the design of professional development programs and instructional resources. Moreover, recent research has demonstrated the potential for integrating insights from CTA into the development of LLMs to generate educational materials that align with expert teachers' practices and preferences. R. E. Wang et al. (2024) found that incorporating expert decision-making models derived from CTA into LLMs led to outputs that were more favorably evaluated by teachers compared to those generated by LLMs without such models. This suggests that LLMs informed by CTA can become valuable tools for generating educational outputs that resonate with educators and reflect best practices in instructional design.

3 Cognitive Task Analysis for Curriculum Scaffolding

To systematically uncover how expert middle school math teachers scaffold their curriculum to meet the diverse needs of their students (RQ1), we conducted a CTA with six expert teachers. The CTA allowed us to formalize teachers' curriculum scaffolding as a three-stage process, which we describe in Section 3.2.

3.1 Methods

Participants. We recruited six middle school math teachers from two public school districts in Washington State and Illinois. Each teacher was selected by district administrators based on their extensive teaching experience (a minimum of 10 years) and perceived expertise, a method that has been shown to effectively identify highly skilled teachers (Jacob & Lefgren, 2008). Collectively, the participants had experience working in a range of public schools and student contexts, ensuring a diverse set of perspectives. All teachers who participated in this phase of the study were compensated for their time at a rate of $45 per hour. The number of experts included in this study is comparable to other natural language processing studies that have worked closely with domain experts (Sharma et al., 2023; R. E. Wang et al., 2024).

Data collection. Data collection consisted of two main components: weekly surveys and CTA interviews. Each week, the expert teachers completed a short survey in which they provided (a) the official (Common Core-aligned, district-designated) curriculum materials for the lessons they were teaching that week, (b) the teacher-intended curriculum materials they specifically prepared to use in their classrooms, and (c) brief descriptions of the changes they had made to the original materials. This approach allowed us to capture the transition from the official curriculum to the teacher-intended curriculum, shedding light on the adaptations and scaffolds the teachers had created.

Prior to each CTA interview, the research team reviewed the official and teacher-intended curriculum materials shared by the teachers and formulated hypotheses about the adaptations made. The interviews were then structured around three main questions: (a) What student responses to the curriculum did the teachers expect? (b) How did they want to respond to these anticipated responses? (c) Why did they choose to respond in that particular manner? These questions were designed to elicit insight into the complex decision-making processes teachers employ when determining when, how, and why to adapt and scaffold their curriculum materials.

During the interviews, teachers were prompted to provide detailed explanations of their thought processes and reasoning behind the changes they had made to the official curriculum materials. The interviewers used follow-up questions and probes to encourage teachers to elaborate on their responses and to clarify any ambiguities.

Data analysis. We analyzed the data collected through weekly surveys and CTA interviews using a qualitative, iterative approach following the process of thematic analysis (Braun & Clarke, 2006). First, we reviewed the original and adapted curriculum materials to identify patterns in the types of changes made by the expert teachers. We then compared these initial observations against teachers' own descriptions of their adaptations and refined them based on insights gained from the CTA interviews. Next, we coded the interview data using a combination of deductive and inductive coding techniques. Deductive codes were derived from existing literature on curriculum use and scaffolding (Remillard & Heck, 2014) and pedagogical content knowledge (Shulman, 1986), while inductive codes emerged from the data itself, capturing the unique strategies and considerations employed by the expert teachers. We organized the coded data into themes and subthemes representing the key components of the teachers' decision-making processes.

To ensure consistent coding and interpretation, our four-member research team held regular discussions to resolve discrepancies and reach consensus. We refined the coding scheme iteratively to accurately represent the data. To enhance trustworthiness, we documented the coding process and decisions, providing transparency in interpretations and theme development. The final themes and subthemes were used to construct a framework describing expert teachers' curriculum scaffolding approach.

Limitations. The small sample size (N=6) and the focus on two specific school districts may limit the generalizability of the findings to other contexts. Additionally, reliance on self-reported data from the teachers may have introduced some degree of bias (Montibeller & Von Winterfeldt, 2015).

3.2 Findings

Based on our findings from the CTA, and building on existing frameworks of curriculum use (Brown, 2002; Sherin & Drake, 2009), we propose a framework for understanding how expert teachers scaffold curriculum materials to meet the needs of students who are struggling to access and engage with the content. This framework comprises three key stages: observation, strategy formulation, and implementation.

3.2.1 Observation: Assessment of Curriculum Materials and Student Needs

The curriculum scaffolding process begins with teachers making observations about the existing curriculum materials and assessing how well they align with their professional intentions and the needs of their students. Consistent with prior research (Collopy, 2003; Sherin & Drake, 2009), our CTA revealed that teachers often read curriculum materials with their students in mind, making determinations about how to use and adapt suggested activities based on their perceptions of students' needs and deficits.

The most common observation made by the expert teachers in our study was the significant gap between the curriculum materials and the current level of their students, particularly among those in low-income communities with high percentages of students working below grade level. One teacher who highlighted this issue explained, “A lot of our students are coming in, not at grade level. They’re coming in on the average anywhere from two to three grade levels below where we are expected to ... get them to within one calendar school year” (Teacher 2). Teachers were acutely aware of this gap and expressed concerns about students’ ability to engage with the materials as presented, with one participant noting that this misalignment could lead to disengagement and behavioral issues.

In addition to concerns about the skill level of students, teachers identified issues with the curriculum’s presentation or the amount of content it covered. For instance, one teacher observed, “The other thing that I noticed about this lesson, it was too much to do all at once” (Teacher 1). Such concerns often intersected with teachers’ professional and pedagogical beliefs about effective instruction. In cases where the curriculum relied heavily on open-ended questions and student discussions, teachers noted that they needed to incorporate more opportunities for direct instruction to support student learning. One teacher shared how students benefited from concrete, visual representations: “It’s great for students who are struggling with the concept of what area means ... because they can see the amount of squares that the shaded part covers. It’s also really good for my high-achieving students because they understand the application” (Teacher 1).

These observations were not only shaped by teachers’ professional judgments but also by their deep understanding of their students’ unique needs. One teacher explained that certain cultural references in the curriculum might not be relevant or accessible to all students, particularly in classrooms with high numbers of English language learners or students with IEPs: “I do have quite a few English language learners, and we don’t always have the resources available to help those students. ... I have about four students who have an IEP for language arts in reading and writing” (Teacher 2).

Moreover, school-level dynamics and district priorities also played a role in shaping how teachers observed and responded to the curriculum materials. Some mentioned that efforts to increase formative assessment or to incorporate more culturally relevant teaching practices were not always reflected in the curriculum. These broader institutional goals often influenced their decisions about how to adapt the materials for their students.

3.2.2 Strategy Formulation: Developing Approaches to Address Observations

After making observations about the curriculum materials, teachers formulated strategies to address each of these observations. These strategies were typically developed at the lesson level and reflected teachers' professional experience and knowledge of what had worked well for them in the past. In particular, our CTA identified several common strategies employed by expert teachers that were grounded in learning sciences and supported by research on effective and accessible instructional practices (CAST, 2024; Institute of Education Sciences, 2021):

- **Activating and refreshing students' background knowledge:** Teachers often began by ensuring that students had the foundational knowledge necessary to engage with new content. For instance, if a lesson on fractions was planned, a teacher might include a task or activity that reviewed basic concepts of division and multiplication. This strategy ensured that students were prepared to engage with the upcoming material. As one teacher explained, "We had to go back to ... solving for unknown variables. We went through a mini lesson [and] did a bunch of practice problems just to refresh our memory on how that process works and what we need to do" (Teacher 3).

- **Supporting the decoding of text, mathematical notation, and symbols:** Teachers frequently added glossaries or visual aids to help students understand complex terminology or symbols, making the content more accessible.

- **Incorporating additional formative assessment:** Formative assessments, such as quick quizzes or exit tickets, helped teachers gauge student understanding throughout the lesson and make real-time adjustments. These assessments served as checkpoints to identify areas where students may have needed additional support before moving forward.

- **Explicitly modeling necessary skills and procedures:** When students struggled with solving equations or other challenging tasks, teachers provided step-by-step demonstrations and think-alouds to model the problem-solving processes. This strategy helped clarify complex tasks and ensured students understood the procedures involved. One teacher explained, "[I provide] upwards of four to six practice problems for them to see that same process over and over, trying to be mindful of my vocabulary, using the same terms and phrases, and using that same repetition, so as they go through it, they can hear me" (Teacher 3). By repeating key processes and language, teachers helped to ensure that students internalized these skills.

- **Providing opportunities for additional practice:** To reinforce key skills and concepts, teachers often described including fluency drills or extra practice problems.

- **Guiding student information processing:** Teachers frequently created structured notetaking guides or graphic organizers to help students capture and organize information effectively. For example, one teacher noted, "I will create for them a notetaking document. ... I will condense [the PowerPoints] into one along with any vocabulary that they need to be familiar with for that particular lesson" (Teacher 2). By providing these structured supports, teachers helped students manage the cognitive load and focus on the essential content.

3.2.3 Implementation

The final stage of the curriculum scaffolding process involved implementing the strategies developed in the previous stage. This implementation occurred primarily at the resource level, with teachers describing two main approaches: adaptation and supplementation of existing curriculum materials.

Adaptation often involved light-touch edits and modifications to the original materials, such as adding definitions, reducing cognitive demand, and making other minor tweaks to improve accessibility and relevance for students. For instance, one teacher described how they reformatted a task with a visual aid to support students: “I set up a hand-drawn coordinate grid to show some of the problems ... and guided students through it ... getting them to provide me with the information” (Teacher 3). This was a relatively light-touch change to how content was presented in the curriculum. Adaptations of this kind were more common when the original curriculum resources aligned with the teacher’s strategy, allowing them to modify the materials in a way that better suited their students’ needs.

Supplementation, on the other hand, usually occurred in response to identified gaps or missing elements in the original materials. This often included developing warmup tasks to review relevant background knowledge, crafting formative assessments to check for understanding, or creating notetaking guides to support student learning. One teacher explained their approach to helping students process information: “I like to provide the notetaking document so the students have a means to process information. So if they’re seeing it, and they’re hearing it, and they’re writing it ... they can interact with those notes” (Teacher 2). This process allowed students to engage with the material through multiple modalities, reinforcing their understanding.

Supplementation frequently involved drawing on third-party repositories, such as Teachers Pay Teachers, where teachers could find suitable materials to incorporate into their lessons. By supplementing the official curriculum, teachers ensured that their students had the necessary supports to succeed, even when they felt the original materials were insufficient.

Figure 1 summarizes the most common observations made by the experts, the associated strategies they employed, and illustrative examples of how they implemented each strategy in their lessons. We note that multiple strategies can be implemented via a single classroom scaffold. For example, a warmup task could be used to activate relevant background knowledge and to build student understanding of important language that will be used in

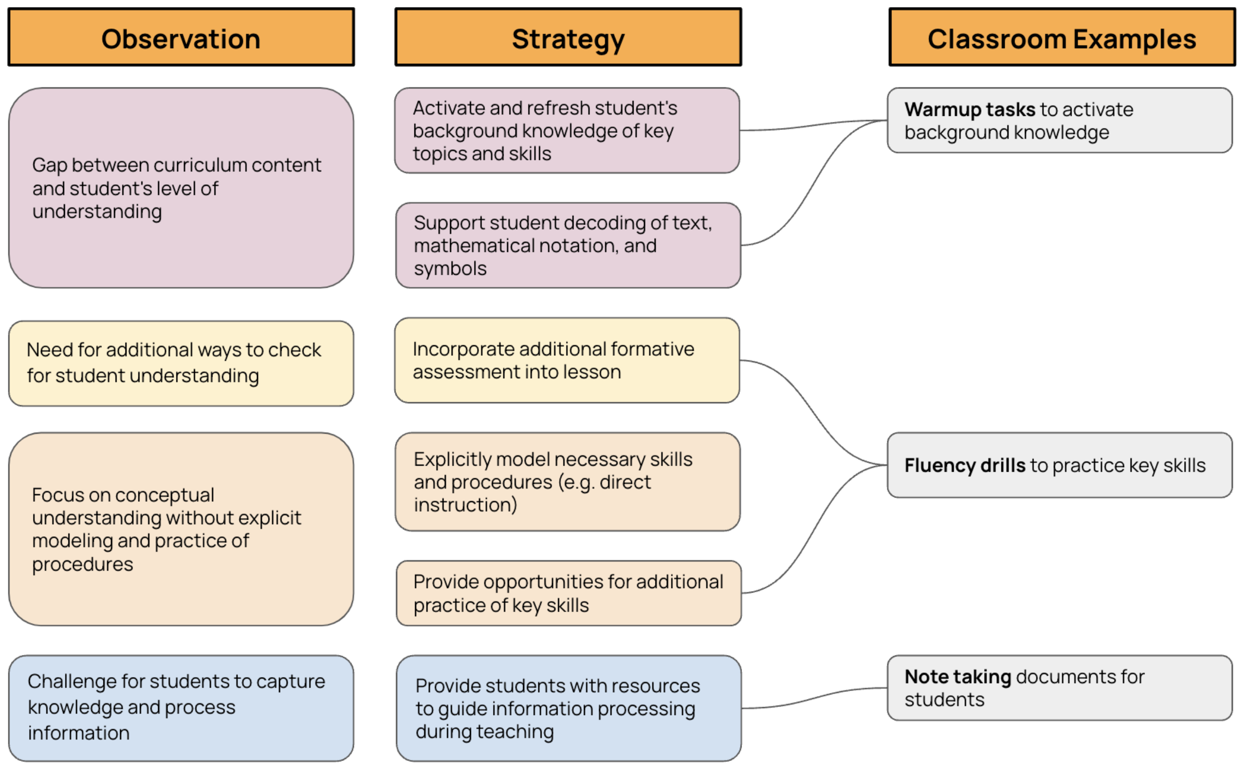


Figure 1: Framework showing the stages of curriculum scaffolding found in the CTA. Expert teachers described first making observations about the existing curriculum materials, then crafting strategies to address these observations, and finally implementing these strategies through adaptations or modifications at the resource level. LLMs can potentially support teachers at each of these steps.

the lesson. Appendix A provides additional examples of curriculum scaffolding strategies and implementation approaches.

3.3 Opportunities for AI Augmentation

The findings from our CTA highlight several aspects of the curriculum scaffolding process that could potentially be augmented through AI approaches. At each stage of the framework, we see opportunities for LLMs to support teachers in their efforts to create accessible and engaging instructional materials.

At the observation stage, LLMs could assist teachers by analyzing curriculum materials to identify potential gaps and areas of challenge. Similar approaches have been used to identify student misconceptions in programming (Fwa, 2024) and math education (Smart et al., 2024). By providing teachers with preliminary observations based on student performance data and patterns in student learning over time, LLMs could save teachers time and offer a foundation upon which to build their instructional strategies.

At the strategy formulation stage, LLMs could be designed to suggest strategies that address the observations made by teachers. For example, if a gap in background knowledge is identified, the LLM could propose strategies to activate prior knowledge. These suggestions could be grounded in learning sciences and professional development best practices, ensuring pedagogical soundness. LLMs could also potentially personalize strategy recommendations based on specific student profiles and learning needs, similar to how they have been used to interpret and explain students' wrong answers (Smart et al., 2024).

At the implementation stage, LLMs could generate scaffolded instructional materials, such as modified tasks or supplementary worksheets, aligned with the teacher's desired strategies. By incorporating expert decision-making models, LLMs could provide suggestions that mimic the nuanced adaptations and supplements typically employed by expert teachers. For example, LLMs could create warmup tasks that review relevant prior content or generate formative assessment questions tailored to the lesson objectives. This aligns with recent work in which LLMs have been used to generate a wide range of educational materials, as outlined in Section 2.3.

These opportunities for AI augmentation resonate with Brown (2002)'s suggestions for supporting teachers' pedagogical design capacity. Brown proposes that teachers should receive help in evaluating the features and affordances of curriculum materials and identifying necessary modifications to align these materials with instructional goals. The use of LLMs to provide this kind of support could not only save teachers time and effort but also potentially serve as a form of embedded professional development, helping teachers to refine their understanding of instruction and student learning.

4 Case Study: Using LLMs to Generate Curriculum Scaffolds

Building on the three-stage framework obtained through our CTA, we conducted a case study to understand how LLMs might be used at the *implementation stage* of curriculum scaffolding, and how these scaffolds compare to those created by experts (RQ2). We restrict our case study to only one stage in order to develop a proof-of-concept with a simpler scenario that does not require jointly evaluating modeling decisions for all three stages at once. Among the three stages, we chose to prioritize implementation because this tends to be the most time-intensive step for teachers (Grossman & Thompson, 2008), and because this stage produces observable outputs that are directly comparable to expert implementations.

For this case study, we drew on a predefined *observation* that a significant proportion of students find the designated curriculum inaccessible, creating a gap between curriculum content and students' current understanding, as supported by teacher interviews and literature (Remillard, 1999; Remillard & Heck, 2014). To address this gap, we employed a scaffolding *strategy* identified during the CTA: activating and refreshing students' background knowledge of key topics and skills. This strategy aligns with research in learning science, which emphasizes the importance of linking new content to prior knowledge to enhance understanding and retention (Dochy et al., 1999). For example, the Universal Design for Learning framework and the National Council of Teachers of Mathematics (NCTM) emphasize building on students' prior knowledge and experiences (CAST, 2024; National Council of Teachers of Mathematics, 2016). However, implementing this in diverse classroom contexts is challenging, particularly given the variation in student backgrounds. By choosing this strategy, we seek to support teachers in making curricula more accessible by helping them activate and connect students' relevant prior knowledge. We leave the task of evaluating LLM performance with respect to producing observations and scaffolding strategies for future work.

Given their theoretical backing and prevalence in teaching, warmup tasks were chosen for this case study to compare LLM-generated scaffolds with those created by expert teachers. Warmup tasks are an effective method for activating background knowledge, providing a brief review of previously learned skills to help students—particularly those learning below grade level—engage with the new lesson material. Research shows that warmup tasks bridge the gap between students' current knowledge and curriculum demands by making new content more accessible (Bransford et al., 2000). For example, a warmup for a lesson on the Pythagorean theorem might review square numbers and right-angled triangles, preparing students to connect foundational concepts to new material. These types of tasks provide students with a structured entry into new content and offer an ideal context to examine how LLM-generated scaffolds compare to expert-created ones, specifically in the implementation stage of scaffolding.

4.1 Methods

4.1.1 Data Collection

Creating a comparison set for LLM outputs required us to collect example implementations of the teacher-intended curriculum. As noted in Section 2.1, this data is challenging to collect, as it is not centrally stored and often exists within individual teacher minds, files, and presentations. As such, we collaborated with two expert teachers from school districts in Washington State and Illinois to collect this novel dataset. Each expert had a minimum of 10 years of teaching experience, had deep familiarity with the Illustrative Mathematics curriculum, and had roles as subject and curriculum leaders within their schools where they mentored other teachers. These qualifications helped to ensure that the dataset was both robust and reflective of high-quality instructional practices. The teachers were compensated at a rate of $50 per hour for their contributions.

The data collection process began with selection of 10 lesson plans from Units 2–7 of the Illustrative Mathematics 6th-grade curriculum. These units target common misconceptions in algebra, categorized into four primary areas: ratios and proportional relationships, the number system, expressions and equations, and functions (Bush & Karp, 2013). The lessons were evenly distributed across these categories, covering 21 of the 26 unique standards within these units. This selection was deliberate, focusing on critical areas of need and spanning a broad range of skills. It is noteworthy that while there is a warmup in every Illustrative Mathematics lesson, the CTA revealed that they often assumed students were on grade level and sometimes did not effectively activate background knowledge. Therefore, modification was necessary to better meet the needs of students working below grade level.

Each expert was provided with the context of these 10 lesson plans, which included comprehensive details, such as the lesson narrative, learning goals, standards, and specific instructional routines and activities. Importantly, the experts were instructed to assume they were teaching the lesson to a class where 50% of the students were working below grade level. They were asked to design a warmup task that would activate background knowledge, that was concise enough to fit on a single Google Slide, and that would align with common practices for creating instructional materials. Additionally, they were asked to provide a brief commentary on the rationale for their created task, including how, if at all, they used the original curriculum warmup in their version. This commentary captured each teacher's resource-level decision-making process.

The dataset comprised 20 items in total, with each lesson plan having two modified warmup tasks (one from each expert). Figure 2 illustrates an example of a lesson plan context, the expert-created warmup task, and the commentary provided by the teachers.

4.1.2 Model Development

To generate the LLM-based warmup tasks, we used GPT-4o, which was selected for its robust performance in natural language processing.[1] Recent literature on LLMs highlights concerns regarding hallucination instances where the model generates content that is incorrect, irrel-

[1]The model was queried in May 2024.

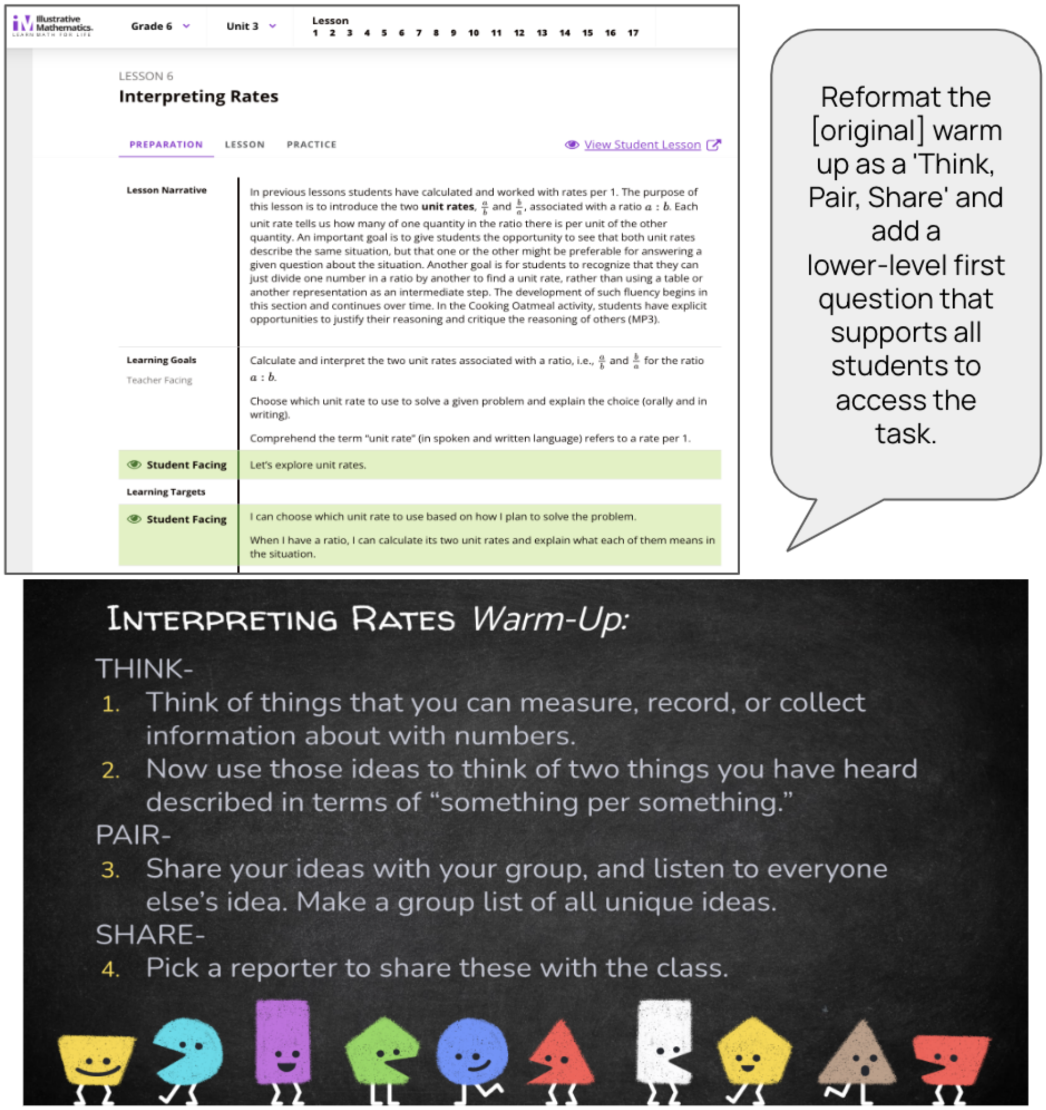


Figure 2: For each lesson context (top left), the expert teachers were given a predefined observation (curriculum inaccessibility) and scaffolding strategy (activating background knowledge). They implemented this strategy by creating a student-facing warmup resource (bottom) and provided commentary on their rationale and use of the original curriculum warmup, capturing their resource-level decision-making process (top right).

evant, or unsupported by the input data (Ji et al., 2023b). In educational contexts, this could manifest as fabricating incorrect mathematical concepts, misrepresenting standards, or suggesting nonexistent instructional strategies. These issues are particularly relevant in our setting, where maintaining the accuracy and educational relevance of generated tasks was critical. However, we did not encounter any hallucinations during our study, likely due to the fact that we only asked it to generate mathematical questions, not answers. By manually verifying each generated warmup task, we confirmed that all outputs were aligned with the curriculum and free of fabricated content. This focused scope, combined with careful manual review, significantly reduced the chances of hallucinations appearing in the scaffolds included in the evaluation.

We prompted the model for each lesson (N=10) based on four prompting strategies, which included zero-shot and three approaches described below. We used the GPT-4o API to generate outputs, allowing us to compare results across these strategies, all of which were informed by findings from the CTA:

1. **Expert-Informed Prompt:** In this condition, each model was provided with the learning goals and lesson narrative. The prompt included general best practices for creating effective warmups, incorporating insights gathered from the CTA. For example, the model was encouraged to consider the skills and knowledge that would be important for that particular lesson, mimicking the thought process that expert teachers described.

2. **Additional Curriculum Context:** In addition to the learning goals and lesson narrative provided in the Expert-Informed Prompt condition, the model was given the original warmup task from the Illustrative Mathematics curriculum. This condition tested the model's ability to utilize existing curriculum components to generate a related but distinct educational task.

3. **Expert Implementation Guidance Provided:** As detailed in Section 4.1.1, the expert teachers involved in data collection provided commentary focused on the implementation stage, justifying how and why they either adapted or supplemented the original warmup in the curriculum, in line with the process outlined in Section 3.2.3. In this third condition, the model received the original warmup task and the expert teacher's commentary on implementation. This condition tested the model's capability

to integrate human expert insights to refine or alter educational content meaningfully. The rationale for this condition was to explore whether providing the model with explicit guidance on how to implement the scaffolding strategy, as opposed to just the original task itself, would lead to higher-quality outputs.

For reference, the full prompts used for each condition are documented in Appendix C.

4.1.3 Evaluation Setup

Our evaluation methodology employed the comparative judgment technique, widely recognized for its reliability in educational research and natural language processing (Pollitt, 2012). To ensure consistency and reliability in evaluator judgments, we implemented strategies such as randomizing lesson order and comparison conditions. For construct validity, we aligned evaluation criteria with established teaching frameworks, ensuring that the aspects that were measured genuinely reflected instructional quality and student engagement. We recruited a diverse group of 50 participants, 75% of whom were current math teachers with varying levels of experience. This diversity enhances the relevance of our findings to the target population of educators, increasing external validity and generalizability.

The evaluations focused on four key criteria: alignment to learning objectives, accessibility for below-grade-level students, readiness for classroom use, and evaluator preference. These criteria were selected for their importance to effective classroom practices and high-leverage instructional goals in mathematics education (CAST, 2024). The focus on accessibility for below-grade-level students addresses common concerns regarding the gap between curriculum content and students' current knowledge levels.

We also recognized the potential influence of differing presentation styles on evaluators' perceptions. To mitigate this bias, we standardized the presentation of materials using a uniform Google Slides template and theme for both model-generated and expert-created warmups. We performed minimal postprocessing on the model outputs: We formatted the text without altering its content and added up to one image per slide to ensure aesthetic consistency. These images, chosen to match the context of the tasks (e.g., an image of paints for questions involving paint), served only aesthetic purposes, keeping the evaluators' focus on content rather than presentation. Figure 3 illustrates the comparison screen shown to evaluators.

Evaluators. To facilitate a comprehensive assessment, we engaged evaluators from two distinct sources: a group of 19 current or former middle school math teachers recruited through partnerships with two school districts in Washington State and Illinois and through personal and institutional networks, as well as an additional 31 active teachers recruited through the Prolific platform. Prolific participants were screened to ensure they were based in the United States or the United Kingdom, were current teachers, had at least an undergraduate degree, and had their degree in mathematics or statistics. Although it was not possible to filter directly for math teachers, we assumed it was likely they had experience teaching math if they met these criteria.

Table 1 summarizes the sample of evaluators based on self-reported information we collected at the beginning of the task. The sample consisted primarily of current math teachers, with 74% actively teaching math. An additional 18% were former math teachers, while the remaining 8% were teachers of another subject with a math degree. In terms of teaching experience, 52% had more than 10 years of experience, 28% had 5–10 years, and 20% had less than 5 years. The majority of evaluators taught at the high school level (78%), followed by middle school teachers (61%); a smaller proportion taught at the elementary level (13%). The sum of the percentages for teacher grade levels exceeds 100% because some reported teaching more than one grade level.

Evaluation task. Each evaluator completed 10 paired comparisons, one for each of 10 distinct lessons. These sixth-grade mathematics lessons were derived from our dataset (Sec-

Table 1: Demographic Information of Evaluators

Category	**Percentage**
Current Math Teacher	74%
Former Math Teacher	18%
Never Taught Math	8%
More than 10 years of teaching experience	52%
5 - 10 years of teaching experience	28%
Less than 5 years of teaching experience	20%
High School Level	78%
Middle School Level	61%
Elementary Level	13%

tion 4.1.1). In each comparison, they evaluated the expert-made warmup against one of four others: the original curriculum warmup and three LLM generated warmups (one from the Expert-Informed Prompt condition, one from the Additional Curriculum Context condition, and one from the Expert Implementation Guidance Ccondition). The original curriculum warmup was included as a baseline comparison, to assess whether evaluators preferred the unmodified materials from Illustrative Mathematics to the versions modified by the expert teacher. In addition to making their selection, evaluators were required to provide two to three sentences justifying their choice. To mitigate potential biases, we randomized the order of the lessons, the selection of the expert warmup (Expert 1 or Expert 2) for each lesson, and the comparison condition presented for each lesson.[2]

In total, 500 comparisons were conducted—125 for the Expert-Informed Prompt condition, 132 for the Additional Curriculum Context condition, 122 for the Expert Implementation Guidance condition, and 121 for the Original Curriculum Warmup condition.

Evaluators assessed the warmups based on the following criteria, chosen for their relevance to practical and effective classroom instruction:

1. **Alignment to Learning Objectives:** Evaluators rated each scaffold based on its

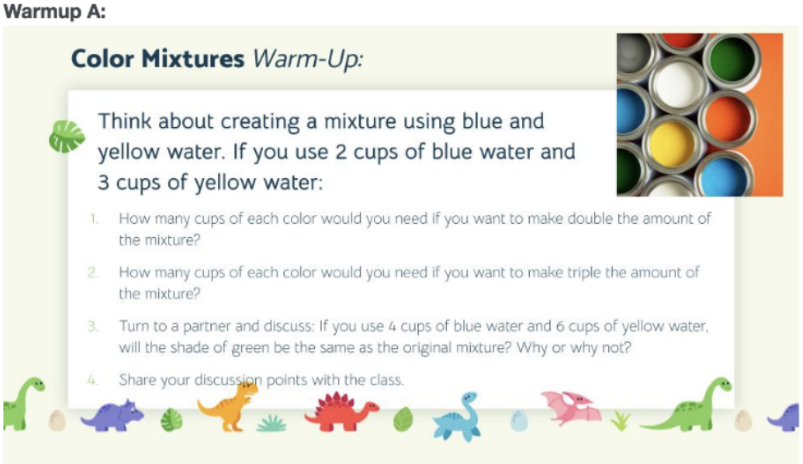
[2] We verified that randomization worked correctly, as each unique model/original warmup item was compared 6.25 times with a standard deviation of 1.196, which is in line with expectations.

You are a 6th grade math teacher selecting a warmup to prepare students for the following lesson. Use the lesson title and learning objectives below to answer the questions below.

Lesson Title: Color Mixtures
Learning Objectives:

- Comprehend and respond (orally and in writing) to questions asking whether two ratios are equivalent, in the context of color mixtures.
- Draw and label a discrete diagram with circled groups to represent multiple batches of a color mixture.
- Explain equivalent ratios (orally and in writing) in terms of the amounts of each color in a mixture being multiplied by the same number to create another mixture that is the same shade.

Warmup A:



Warmup B:

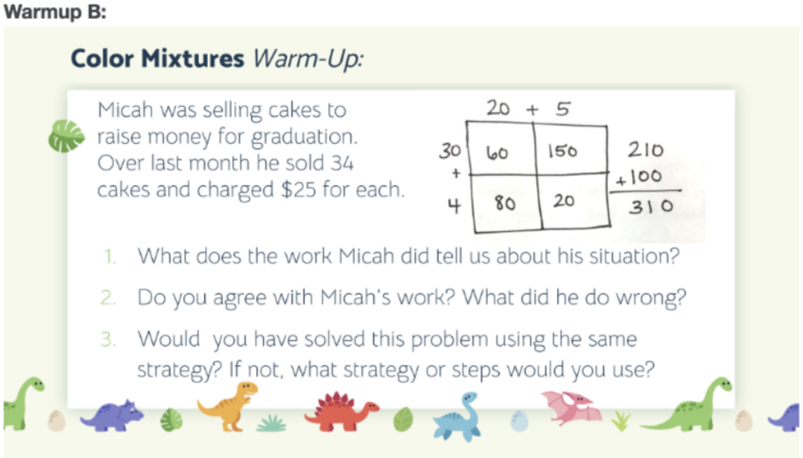


Figure 3: For each of the 10 lessons, evaluators were shown two warmups, one created by an expert and one created in one of our model conditions. The order of expert and model were randomized each time.

alignment to the specific learning objectives, ensuring that the task directly contributed to achieving these goals.

2. **Accessibility for Below-Level Students:** This criterion examined the ease with which below-level students could engage with the task, considering the clarity, simplicity, and support embedded within the scaffold.

3. **Readiness for Classroom Use:** This measured the extent to which a scaffold could be utilized in the classroom without further modification. Evaluators assessed how much additional preparation, adaptation, or modification each scaffold would require before it could be used effectively. Readiness for use is important, as it reduces preparation time, allowing teachers to focus more on instruction and interaction with students.

4. **Preference:** Evaluators were asked which warmup they would prefer to use if they were teaching the specific learning objective. This captured the overall practical preference of educators based on their professional judgment.

Each of the dimensions listed above was assessed using a direct comparison between the expert-created and model-generated warmups. Evaluators indicated their preference on a 5-point Likert scale ranging from "Warmup A is much more [aligned to learning objectives/accessible/ready-to-go/preferred]" to "Warmup A and B are equally [aligned to learning objectives/accessible/ready-to-go/preferred]" to "Warmup B is much more [aligned to learning objectives/accessible/ready-to-go/preferred]." These ratings were then coded on a scale from -2 to +2, with -2 and -1 corresponding to Warmup A being much more or somewhat more preferred, 0 indicating equal preference, and +1 and +2 corresponding to Warmup B being somewhat more or much more preferred.

4.2 Results

Table 2 summarizes the results of the comparative judgments between the model-generated warmups and the expert-created benchmarks across four dimensions: alignment to learning objectives, accessibility for below-level students, readiness for classroom use, and preference.

Metric	Expert-Informed Prompt	Additional Curriculum Context	Implementation Guidance	Original Curriculum Warmup
...aligned with learning objectives.	0.528***	0.636***	0.467***	-0.331**
...accessible.	0.448***	0.447***	0.410**	-0.099
...ready-to-go.	0.440***	0.371***	0.270*	0.017
...preferred.	0.496***	0.545***	0.410**	-0.124

Table 2: Results of Comparative Judgments Between Model-Generated Warmups and Expert-Created Benchmarks

The results are also visually represented in Appendix B, which shows violin plots for each of the dimensions assessed and a brief discussion of normality and independence.

Overall, the Additional Curriculum Context condition performed best across all evaluation criteria. These outputs were significantly preferred to the expert outputs in terms of alignment to learning objectives ($M = 0.636$, $p < 0.001$), accessibility ($M = 0.447$, $p < 0.001$), readiness for classroom use ($M = 0.371$, $p < 0.001$), and overall preference ($M = 0.545$, $p < 0.001$). However, the magnitude of the difference is moderate, with an average of 0.50 across the criteria on the -2 to 2 point scale.

The Expert-Informed Prompt condition also performed well, with outputs significantly preferred to the expert versions in alignment to learning objectives ($M = 0.528$, $p < 0.001$), accessibility ($M = 0.448$, $p < 0.001$), readiness for classroom use ($M = 0.440$, $p < 0.001$), and overall preference ($M = 0.496$, $p < 0.001$). The magnitude of the difference is similar to the Additional Curriculum Context condition. A paired t-test indicates that the scores for the Expert-Informed Prompt condition are not significantly different from the Additional Curriculum Context condition for alignment to learning objectives ($t = -0.649$, $p = 0.517$), accessibility ($t = -0.238$, $p = 0.812$), readiness for classroom use ($t = 0.163$, $p = 0.871$), and overall preference ($t = -0.342$, $p = 0.733$).

The Implementation Guidance Provided condition also performed well, with outputs significantly preferred to the expert versions in terms of alignment to learning objectives ($M = 0.467$, $p < 0.001$), accessibility ($M = 0.410$, $p < 0.01$), readiness for classroom use ($M = 0.270$, $p < 0.05$), and overall preference ($M = 0.410$, $p < 0.001$). A paired t-test indicates that the scores for the Implementation Guidance Provided condition are not significantly different from the top-performing Additional Curriculum Context condition for alignment to learning objectives ($t = 1.051$, $p = 0.295$), accessibility ($t = 0.533$, $p = 0.595$), readiness for classroom use ($t = 1.063$, $p = 0.290$), and overall preference ($t = 1.037$, $p = 0.302$). We provide hypotheses for why this condition did not outperform the other conditions in the discussion and future work sections.

The Original Curriculum Warmup condition was not significantly preferred to the expert outputs on most of the evaluation criteria, with mean scores close to zero for accessibility ($M = -0.099$, $p > 0.05$), readiness for classroom use ($M = 0.017$, $p > 0.05$), and overall preference ($M = -0.124$, $p > 0.05$). However, for alignment to learning objectives, the expert-created warmups were significantly preferred to the original warmups in the curriculum ($M = -0.331$, $p < 0.01$). This suggests that while the original curriculum warmups were perceived as comparable to the expert-created ones in terms of accessibility, readiness for use, and overall preference, the expert-created warmups were considered better aligned with the learning objectives of the lessons.

5 Discussion

Our research aimed to explore the potential of LLMs in supporting teachers with scaffolding their curricula, specifically focusing on middle school mathematics. We were particularly interested in the transition from the designated curriculum to the teacher-intended curriculum and, within that, the subset of adaptations and supplementations that teachers employ to ensure that all students can access and engage with the curriculum, which we define as curriculum scaffolding.

Building on CTA with expert teachers, we proposed a three-step process for curriculum scaffolding: observation, strategy formulation, and implementation. This framework aligns with existing literature, such as Sherin & Drake (2009)'s conceptualization of teachers reading, evaluating, and adapting the curriculum and Brown (2002)'s three modes of engagement with curriculum—offloading, adapting, and improvising. We extend prior work by providing additional granularity specific to the context of preparing curriculum for middle school mathematics classrooms where students are struggling to access and engage with the content. Our framework offers insights into the day-to-day implementations of various adaptation strategies in this specific subject and context, some of which are described in Appendix A. For example, we provide additional color to what "adapting" means in Sherin & Drake (2009)'s conceptualization, detailing specific techniques such as activating background knowledge, supporting the decoding of text and symbols, and explicitly modeling necessary skills and procedures. These insights contribute to a more nuanced understanding of curriculum scaffolding in contexts of learner variability.

To apply this framework and investigate the potential of LLMs to produce high-quality curriculum scaffolds, we focused on a specific scenario and strategy identified in our CTA with expert teachers: creating warmup tasks that activate and refresh students' background knowledge. We presented a novel dataset of expert-created warmup tasks for this purpose and developed several LLM-based approaches informed by the insights gained from the CTA. Our evaluation involved 500 comparisons across three model conditions and the original curriculum warmup by math teachers, and we found that the LLM-generated warmup tasks performed better than those created by expert teachers. The best-performing approach was the Additional Curriculum Context condition, where the model was provided with the original warmup task from the curriculum in addition to an expert-informed prompt. This finding resonates with the expert teacher approach identified in the CTA, wherein teachers review the original curriculum materials before deciding how to adapt them for their specific context.

Our results highlight the potential for LLMs to support teachers in the curriculum scaffolding process, particularly when provided with relevant context and guidance informed by expert practices. By automating the creation of high-quality instructional scaffolds—such as warmup tasks—LLMs could help alleviate the burden on teachers, allowing them to focus more on direct instruction and interaction with students. Furthermore, our work demonstrates an approach for closely involving teachers in the development and evaluation of LLM-based approaches for education, which the research community has called for (Nazaretsky et al., 2022). By replicating this approach for other specific curriculum scaffolding strategies and implementation examples, researchers can continue to explore the potential of AI to support teachers in creating effective, engaging, and inclusive learning experiences for all students.

Our findings contribute to the growing body of literature on curriculum use and the application of AI in education. By providing a more granular understanding of the curriculum scaffolding process and demonstrating the potential of LLMs to generate high-quality instructional materials, this study lays the groundwork for the development of AI-driven tools that can support teachers in their efforts to meet the diverse needs of their students, particularly in contexts where many learners are struggling to access and engage with the curriculum.

6 Limitations and Future Work

While our study provides valuable insights, it is only a first step towards understanding the potential of LLMs to support curriculum scaffolding.

One limitation is the reliance on a small set of expert-created warmups for evaluation. Although we collaborated with experienced teachers to develop the dataset, having only two examples of teacher-created warmups per lesson may not fully capture the diverse ways in which warmups can be implemented. Evaluators' preferences might have been influenced by factors beyond the quality of the warmups themselves. To address this limitation, future studies should aim to expand the dataset to include a broader range of expert-created warmups, ensuring a more comprehensive and accurate evaluation of LLM-generated scaffolds.

Another limitation is that the strategies used in the Expert Strategy condition were derived from the dataset creators rather than the evaluators themselves. It is possible that the evaluators might have had different ideas about how to improve the warmups, and the strategies infused into the model might not fully reflect their approaches. This discrepancy could have affected the model's performance in this condition. To mitigate this issue, future research should consider a more dynamic approach, where teachers review the original warmup, provide their own strategies for improvement, and then evaluate the automatically generated warmup based on their input. Such an interactive AI-in-the-loop system would help ensure that the LLM-generated scaffolds align more closely with the evaluators' perspectives and preferences.

Our study lays the groundwork for several exciting avenues of research. One priority should be to expand the dataset to include a wider variety of instructional scaffolds beyond warmup tasks, as well as observations and strategies to capture all three stages of the scaffolding process. By applying the framework and methodology developed in this study to other types of curriculum adaptations and supplements, researchers can gain a more comprehensive understanding of the potential for LLMs to support teachers in the curriculum scaffolding process. However, this type of data collection process would need to be dynamic to address the limitations mentioned above.

Additionally, piloting the LLM-generated warmups in real classroom settings would provide valuable insights into their practical effectiveness and help identify areas for further refinement. Collaborative research with teachers, involving the use of AI-generated scaffolds in their day-to-day practice, could shed light on the challenges and opportunities associated with integrating these tools into existing instructional workflows. Moreover, conducting focus groups and interviews with students to qualitatively understand their perception of the materials—and measuring outcomes such as their relationship to math and academic performance—could help assess the impact of these scaffolds on learners, similarly to what other studies on scaffolding have done.

Another promising direction for future research is the exploration of more nuanced and personalized scaffolding strategies. By leveraging real-time feedback from teachers and data on student performance and engagement, LLMs could potentially generate instructional scaffolds that are tailored to the specific needs and characteristics of individual classrooms and learners (Lim et al., 2024). The development of adaptive, data-driven scaffolding tools could represent a significant step forward in supporting teachers' efforts to create inclusive and effective learning experiences for all students.

7 Ethics Statement

This study was conducted in compliance with ethical research standards. IRB approval was required for the interviews with teachers and their participation in the evaluation activities. IRB approval was obtained in May 2024, and the study was determined to be of "minimal risk."

AI-powered scaffold generation is not without risks: Inappropriate use can exacerbate inequities by compromising student privacy and teacher autonomy, and by perpetuating unequal access to high-quality curricula. In this work, we focused on evaluating LLMs rather than on implementing LLM-generated scaffolds in classrooms. Downstream tools that generate scaffolds intended for teachers' real-world classroom use require extra precaution.

To mitigate data privacy risks, such tools should avoid collecting identifiable information. Teachers should be encouraged to review and adapt generated tasks to suit their specific classroom needs, ensuring both quality control and teacher autonomy. Additionally, to reduce bias in generated outputs, models should be tuned or prompted using expert-designed, high-quality curriculum materials, such as those found in the Illustrative Mathematics curriculum.

In the context of our task, these concerns might arise if the generated scaffolds introduce biased scaffolding strategies or lack cultural relevance, potentially alienating certain student groups or reinforcing existing disparities in education. Our findings emphasize the importance of carefully considering these risks and implementing safeguards to ensure ethical and equitable use of AI in educational settings.

8 Acknowledgments

We are grateful to Stanford HAI and the Stanford Accelerator for Learning for supporting this work.

References

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academy Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi: 10.1191/1478088706qp063oa

Brown, M. W. (2002). *Teaching by design: Understanding the intersection between teacher practice and the design of curricular innovations.* Northwestern University.

Bulathwela, S., Muse, H., & Yilmaz, E. (2023). Scalable educational question generation with pre-trained language models. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 327–339). Cham: Springer Nature Switzerland.

Bush, S. B., & Karp, K. S. (2013). Prerequisite algebra skills and associated misconceptions of middle grade students: A review. *The Journal of Mathematical Behavior*, *32*(3), 613–632. doi: 10.1016/j.jmathb.2013.07.002

CAST. (2024). *Universal design for learning guidelines.* `https://udlguidelines.cast.org/`. (Accessed 15-April-2024)

Clark, R., Feldon, D., Van Merrienboer, J. J. G., Yates, K., & Early, S. (2008, 01). Cognitive task analysis. *Handbook of Research on Educational Communications and Technology*, 577-593.

Collopy, R. (2003). Curriculum materials as a professional development tool: How a mathematics textbook affected two teachers' learning. *The Elementary School Journal*, *103*(3), 287–311.

Diwan, C., Srinivasa, S., Suri, G., Agarwal, S., & Ram, P. (2023). Ai-based learning content generation and learning pathway augmentation to increase learner engagement.

Computers and Education: Artificial Intelligence, *4*, 100110. Retrieved from https://www.sciencedirect.com/science/article/pii/S2666920X22000650 doi: https://doi.org/10.1016/j.caeai.2022.100110

Doan, S., Kaufman, J. H., Woo, A., Tuma, A. P., Diliberti, M. K., & Lee, S. (2022). How states are creating conditions for use of high-quality instructional materials in k–12 classrooms.

Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research*, *69*(2), 145–186. doi: 10.3102/00346543069002145

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., ... Sakr, M. (2024). A comparative study of ai-generated (gpt-4) and human-crafted mcqs in programming education. New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3636243.3636256 doi: 10.1145/3636243.3636256

Drew, C. (2022). *18 scaffolding examples in education.* Helpful Professor. Retrieved from https://helpfulprofessor.com/scaffolding-examples-in-education/ (Accessed: 2024-05-24)

Elkins, S., Kochmar, E., Cheung, J. C. K., & Serban, I. (2023). *How useful are educational questions generated by large language models?*

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, *36*.

Fwa, H. L. (2024). Experience report: Identifying common misconceptions and errors of novice programmers with chatgpt. In *Proceedings of the 46th international conference on software engineering: Software engineering education and training* (p. 233–241). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3639474.3640059 doi: 10.1145/3639474.3640059

Galiatsos, S., Kruse, L., & Whittaker, M. (2019). Forward together: Helping educators unlock the power of students who learn differently. *National Center for Learning Disabilities*.

Gehrke, N. J., Knapp, M. S., & Sirotnik, K. A. (1992). Chapter 2: In search of the school curriculum. *Review of research in education*, *18*(1), 51–110.

Goddard, Y., Goddard, R., & Kim, M. (2015). School instructional climate and student achievement: An examination of group norms for differentiated instruction. *American Journal of Education*, *122*(1), 111–131.

Gonzalez Thompson, A. (1984). The relationship of teachers' conceptions of mathematics and mathematics teaching to instructional practice. *Educational studies in mathematics*, *15*(2), 105–127.

Grossman, P., & Thompson, C. (2008). Learning from curriculum materials: Scaffolds for new teachers? *Teaching and Teacher Education*, *24*(8), 2014-2026. Retrieved from https://www.sciencedirect.com/science/article/pii/S0742051X0800084X doi: https://doi.org/10.1016/j.tate.2008.05.002

Gueudet, G., & Trouche, L. (2009). Towards new documentation systems for mathematics teachers? *Educational studies in mathematics*, *71*, 199–218.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* routledge.

Impact Research. (2024). *Ai chatbots in schools: Findings from a poll of k-12 teachers, students, parents, and college undergraduates.* (Accessed at: https://8ce82b94a8c4fdc3ea6d-b1d233e3bc3cb10858bea65ff05e18f2.ssl.cf2.rackcdn.com/

bf/24/cd3646584af89e7c668c7705a006/deck-impact-analysis-national-schools-tech-tracker-may-2024-1.pdf)

Institute of Education Sciences. (2021). *Assisting students struggling with mathematics: Intervention in the elementary grades.* U.S. Department of Education, What Works Clearinghouse. Retrieved from https://ies.ed.gov/ncee/WWC/Docs/PracticeGuide/WWC2021006-Math-PG.pdf (WWC 2021006)

Irwin, V., Wang, K., Tezil, T., Zhang, J., Filbey, A., Jung, J., ... Parker, S. (2023). *Report on the condition of education 2023. nces 2023-144* (Tech. Rep.). National Center for Education Statistics.

Jackson, K., & Makarin, A. (2018, August). Can online off-the-shelf lessons improve student outcomes? evidence from a field experiment. *American Economic Journal: Economic Policy*, *10*(3), 226–54. Retrieved from https://www.aeaweb.org/articles?id=10.1257/pol.20170211 doi: 10.1257/pol.20170211

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–136. Retrieved 2024-05-24, from http://www.jstor.org/stable/10.1086/522974

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023a). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023b, March). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. Retrieved from http://dx.doi.org/10.1145/3571730 doi: 10.1145/3571730

Jiao, Y., Shridhar, K., Cui, P., Zhou, W., & Sachan, M. (2023). Automatic educational question generation with difficulty level controls. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 476–488). Cham: Springer Nature Switzerland.

Jury, B., Lorusso, A., Leinonen, J., Denny, P., & Luxton-Reilly, A. (2024). Evaluating llm-generated worked examples in an introductory programming course. In *Proceedings of the 26th australasian computing education conference* (p. 77–86). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3636243.3636252 doi: 10.1145/3636243.3636252

Leiker, D., Finnigan, S., Gyllen, A. R., & Cukurova, M. (2023). *Prototyping the use of large language models (llms) for adult learning content creation at scale.*

Lim, L., Bannert, M., van der Graaf, J., Fan, Y., Rakovic, M., Singh, S., . . . Gašević, D. (2024). How do students learn with real-time personalized scaffolds? *British Journal of Educational Technology*, *55*(4), 1309-1327. Retrieved from https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13414 doi: https://doi.org/10.1111/bjet.13414

Lokey-Vega, A. (2015, March). Expert as the tpack misfit: A cognitive task analysis to map expert-teacher cognitive processes of technology-rich lesson planning. In D. Rutledge & D. Slykhuis (Eds.), *Proceedings of society for information technology & teacher education international conference 2015* (pp. 3322–3330). Las Vegas, NV, United States: Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/p/150461

Meidl, T., & Meidl, C. (2011, Apr.). Curriculum integration and adaptation: Individualizing pedagogy for linguistically and culturally diverse students. *Current Issues in Education*, *14*(1). Retrieved from https://cie.asu.edu/ojs/index.php/cieatasu/article/view/579

Montibeller, G., & Von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. *Risk analysis*, *35*(7), 1230–1251.

National Council of Teachers of Mathematics. (2016). *High expectations.* https://www.nctm.org/Standards-and-Positions/Position-Statements/High-Expectations/. ([Online; accessed 15-April-2024])

Nazaretsky, T., Bar, C., Walter, M., & Alexandron, G. (2022). Empowering teachers with ai: Co-designing a learning analytics tool for personalized instruction in the science classroom. In *Lak22: 12th international learning analytics and knowledge conference* (p. 1–12). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3506860.3506861 doi: 10.1145/3506860.3506861

Nerlino, E. (2022, January). Navigating "the chaos": teacher considerations while adapting curriculum and instruction during the COVID-19 pandemic. *Qualitative Research Journal*, *22*(4), 433–447. Retrieved 2024-06-19, from https://doi.org/10.1108/QRJ-02-2022-0026 (Publisher: Emerald Publishing Limited) doi: 10.1108/QRJ-02-2022-0026

Ninaus, M., & Sailer, M. (2022). Closing the loop–the human role in artificial intelligence for education. *Frontiers in psychology*, *13*, 956798.

Norberg, K., Almoubayyed, H., Fancsali, S. E., DeLey, L., Weldon, K., Murphy, A., & Ritter, S. (2023, July). Rewriting math word problems with large language models. In *Aied23: Artificial intelligence in education, empowering education with llms workshop.* Tokyo, Japan.

Philipp, A., & Kunter, M. (2013). How do teachers spend their time? a study on teachers' strategies of selection, optimisation, and compensation over their career cycle. *Teaching and Teacher Education*, *35*, 1–12.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: principles, policy & practice*, *19*(3), 281–300.

Prihar, E., Lee, M., Hopman, M., Kalai, A. T., Vempala, S., Wang, A., ... Heffernan, N. (2023). Comparing different approaches to generating mathematics explanations using large language models. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education. posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 290–295). Cham: Springer Nature Switzerland.

Remillard, J. T. (1999). Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry*, *29*(3), 315-342. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/0362-6784.00130` doi: https://doi.org/10.1111/0362-6784.00130

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of educational research*, *75*(2), 211–246.

Remillard, J. T., & Heck, D. J. (2014). Conceptualizing the curriculum enactment process in mathematics education. *Zdm*, *46*, 705–718.

Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, *208*, 118258.

Sharma, A., Rushton, K., Lin, I., Wadden, D., Lucas, K., Miner, A., ... Althoff, T. (2023, July). Cognitive reframing of negative thoughts through human-language model interaction. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9977–10000). Toronto, Canada: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2023.acl-long.555` doi: 10.18653/v1/2023.acl-long.555

Sherin, M. G., & Drake, C. (2009). Curriculum strategy framework: investigating patterns in teachers' use of a reform-based elementary mathematics curriculum. *Journal of Curriculum Studies*, *41*(4), 467–500.

Shimmei, M., Bier, N., & Matsuda, N. (2023). Machine-generated questions attract instructors when acquainted with learning objectives. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 3–15). Cham: Springer Nature Switzerland.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. doi: 10.3102/0013189X015002004

Smart, F., Bos, N. D., & Bos, J. T. (2024). Can large language models recognize and respond to student misconceptions? In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive instructional systems* (pp. 288–299). Cham: Springer Nature Switzerland.

Squire, K. D., MaKinster, J. G., Barnett, M., Luehmann, A. L., & Barab, S. L. (2003). Designed curriculum and local culture: Acknowledging the primacy of classroom culture. *Science Education*, *87*(4), 468-489. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/sce.10084` doi: https://doi.org/10.1002/sce.10084

Sridhar, P., Doyle, A., Agarwal, A., Bogart, C., Savelka, J., & Sakr, M. (2023). *Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives.*

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics teaching in the middle school*, *3*(4), 268–275.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes.* Harvard university press.

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932–945. doi: 10.1037/a0031882

Wang, R., & Demszky, D. (2023, June). Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *18th workshop on innovative use of nlp for building educational applications.*

Wang, R. E., Zhang, Q., Robinson, C., Loeb, S., & Demszky, D. (2024, July). Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics.* Association for Computational Linguistics.

Wang, Z., Valdez, J., Basu Mallick, D., & Baraniuk, R. G. (2022). Towards human-like educational question generation with large language models. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 153–166). Cham: Springer International Publishing.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100.

Yan, J., & Goh, H. H. (2023). Exploring the cognitive processes in teacher candidates' collaborative task-based lesson planning. *Teaching and Teacher Education*, *136*, 104365.

Zuo, M., Kong, S., Ma, Y., Hu, Y., & Xiao, M. (2023). The effects of using scaffolding in online learning: A meta-analysis. *Education Sciences*, *13*(7), 705.

Appendix A Curriculum Scaffolding Strategies and Implementation Approaches

Curriculum Scaffolding Strategy	Implementation Examples
Increase relevance, value, authenticity UDL Guideline 7.2 (CAST, 2024)	1. Use motivating questions to engage students from the start of the lesson. 2. Introduce the lesson with a hook to capture students' interest. 3. Incorporate students' personal interests and experiences into the lesson content.
Activate background knowledge UDL Guideline 3.1 (CAST, 2024)	1. Start with a task, such as a warmup or quiz, to review and activate prerequisite topics and skills. 2. Provide a handout with relevant background knowledge, including worked examples, helpful formulas, and definitions. 3. Use graphic organizers to help students structure and visualize information. 4. Embed notes within the lesson materials to reinforce key concepts and provide guidance.

Curriculum Scaffolding Strategy	Implementation Examples
Build student fluency with targeted direct instruction and graduated levels of practice. UDL Guideline 5.3 (CAST, 2024)	1. Conduct fluency drills that focus on the specific skill being taught in the lesson. 2. Ask probing questions that vary in difficulty and form to deepen understanding. 3. Use quizzes and drills to reinforce prerequisite skills. 4. Implement digital practice activities that provide immediate feedback to students. 5. Provide direct instruction to address common misconceptions and errors. 6. Model key skills through direct instruction to ensure students understand the process.
Support decoding of text, mathematical notation, and symbols to build math literacy. UDL Guideline 2.3 (CAST, 2024)	1. Create vocabulary lists, word walls, and use strategic word substitutions to support language comprehension. 2. Design practice activities that target students' use of mathematical language in both large group and small group tasks.

Curriculum Scaffolding Strategy	Implementation Examples
Add formative assessment UDL Guideline 9.3 (CAST, 2024)	1. Use exit tickets and diagnostic assessment questions to gauge student understanding at the end of lessons. 2. Implement think, pair, share activities to encourage student discussion and reflection. 3. Conduct low-stakes quizzes to monitor ongoing progress without adding pressure. 4. Provide digital practice with immediate feedback to help students correct mistakes in real-time. 5. Include mid-lesson checks for understanding to adjust instruction as needed. 6. Encourage student self-assessment to promote self-reflection and goal setting.
Provide opportunities for deeper conceptual understanding to optimize challenge. UDL Guideline 8.2 (CAST, 2024)	1. Offer more challenging practice that covers the same core skill to push student thinking. 2. Provide practice with immediate feedback to help students self-correct. 3. Design challenge practice activities that build new knowledge based on current skills.

Curriculum Scaffolding Strategy	Implementation Examples
Adapt curriculum to include appropriate goal-setting, planning, and strategy development and managing information to support executive function and SEL. UDL Guideline 6.1 -6.3 (CAST, 2024)	1. Prompt students to plan their problem-solving approach before starting tasks. 2. Encourage students to break down tasks themselves to develop planning skills. 3. Break up complex tasks or lessons into smaller, manageable parts for students. 4. Use flowcharts to guide students through decision-making processes. 5. Provide knowledge organizers with key notes and information to aid in learning.

Curriculum Scaffolding Strategy	Implementation Examples
Adapt curriculum to guide information processing and lighten the cognitive load on learners. UDL Guideline 3.3 (CAST, 2024)	1. Simplify numbers and data in problems to make them more accessible. 2. Use worked examples to demonstrate step-by-step solutions. 3. Include hints and suggestions in handouts, especially guidance for the first step in tasks. 4. Provide lesson notes in handouts to support learning. 5. Break down student-facing material into multiple artifacts, such as warm-ups, in-class practice, individual practice, and exit tickets. 6. Ensure the visual design of modified practice materials is similar to general practice to avoid confusion. 7. Reformat student-facing work to make it more intuitive and easier to follow.

Appendix B Model Evaluation

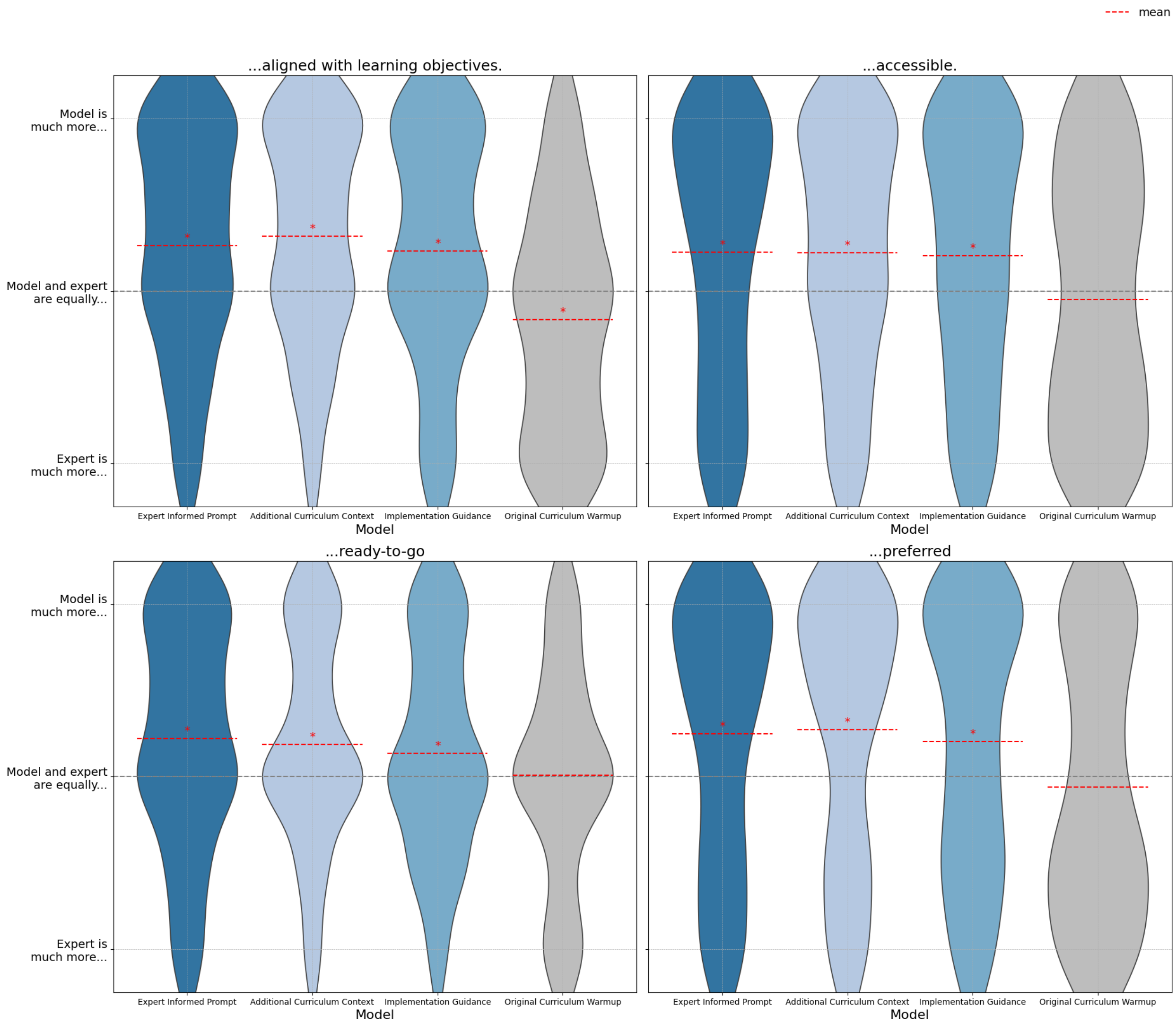


Figure B1: Evaluators compared warmup tasks created under four model conditions (Expert-Informed Prompt, Additional Curriculum Context, Implementation Guidance, Original Curriculum Warmup) to those crafted by experts. The violin plots show the distribution of scores across four dimensions: alignment with learning objectives, accessibility for below-level students, readiness for classroom use, and overall preference. Red dashed lines indicate the mean scores. We checked for normality through visual inspection of the distribution of difference, which appeared approximately normal. The independence assumption was met by design, as each pair of evaluations was independent across participants. These statistical tests confirmed that the differences observed across the conditions were valid for t-test analysis.

Appendix C Model Prompts

Expert-Informed Prompt

```
Your goal is to create a short warmup task that activates and refreshes relevant background knowledge for the lesson information provided.
--------------------
Lesson Title: {row['lesson_title']}
Lesson Narrative: {row['lesson_narrative']}
Learning Objectives: {row['learning_goals']}
--------------------
Class Context: 50% of students are performing below grade level and find it challenging to access the curriculum.
--------------------
Here are tips for creating a great warmup task:
- Think about the specific skills and knowledge students will need to access today's lesson
- If they're provided, you can draw on the original curriculum warmup and teacher response to original warmup
- The warmup should be focused on a single short scenario with up to 3 subparts
- Some subparts should be accessible for all students
- The warmup should take no more than 5 minutes to complete in class
- The task should be student-facing and not need anything else for students to complete it
- Do not provide a title. Do not provide anything apart from the task. Provide response in natural language, do not use latex
- The output should be short, concise and clear. Good warmups usually have 40-100 words
- Think step-by-step
--------------------
Now, create the warmup task.
WARMUP TASK:"""
```

Figure C2: This prompt was iteratively designed with input from expert teachers. It pipes in relevant information from the curriculum.

Expert-Informed Prompt + Additional Curriculum Context

```
Your goal is to create a short warmup task that activates and refreshes relevant background knowledge for the lesson information provided.
--------------------
Lesson Title: {row['lesson_title']}
Lesson Narrative: {row['lesson_narrative']}
Learning Objectives: {row['learning_goals']}
Original Curriculum Warmup: {row['warmup']}
--------------------
Class Context: 50% of students are performing below grade level and find it challenging to access the curriculum.
--------------------
Here are tips for creating a great warmup task:
- Think about the specific skills and knowledge students will need to access today's lesson
- If they're provided, you can draw on the original curriculum warmup and teacher response to original warmup
- The warmup should be focused on a single short scenario with up to 3 subparts
- Some subparts should be accessible for all students
- The warmup should take no more than 5 minutes to complete in class
- The task should be student-facing and not need anything else for students to complete it
- Do not provide a title. Do not provide anything apart from the task. Provide response in natural language, do not use latex
- The output should be short, concise and clear. Good warmups usually have 40-100 words
- Think step-by-step
--------------------
Now, create the warmup task.
WARMUP TASK:"""
```

Figure C3: This prompt included the original curriculum warmup from Illustrative Mathematics.

Expert-Informed Prompt + Additional Curriculum Context + Implementation Guidance

```
Your goal is to create a short warmup task that activates and refreshes relevant background knowledge for the lesson information provided.
--------------------
Lesson Title: {row['lesson_title']}
Lesson Narrative: {row['lesson_narrative']}
Learning Objectives: {row['learning_goals']}
Original Curriculum Warmup: {row['warmup']}
Teacher Response to Original Warmup: {row['teacher_feedback_to_warmup']}
--------------------
Class Context: 50% of students are performing below grade level and find it challenging to access the curriculum.
--------------------
Here are tips for creating a great warmup task:
- Think about the specific skills and knowledge students will need to access today's lesson
- If they're provided, you can draw on the original curriculum warmup and teacher response to original warmup
- The warmup should be focused on a single short scenario with up to 3 subparts
- Some subparts should be accessible for all students
- The warmup should take no more than 5 minutes to complete in class
- The task should be student-facing and not need anything else for students to complete it
- Do not provide a title. Do not provide anything apart from the task. Provide response in natural language, do not use latex
- The output should be short, concise and clear. Good warmups usually have 40-100 words
- Think step-by-step
--------------------
Now, create the warmup task.
WARMUP TASK:"""
```

Figure C4: This prompt also includes the expert teacher's strategy for modification of the original curriculum warmup.

Category	**Details**
LLM Model	GPT-4o (default version, not fine-tuned)
Model Information	Note that the default, zero-shot model was not fine-tuned
API Parameters	Max tokens: 1500 Temperature: 1.0 (default) Top_p: 1 (default) Frequency penalty: 0 (default) Presence penalty: 0 (default)
Model Prompt Strategy	Constructed dynamically based on the row data (e.g., lesson title, objectives, narrative, and feedback). Prompting strategies include zero-shot, curriculum-context-enhanced, and teacher-rationale-enhanced prompts.
Context Window	Default of 32,760 tokens
Date Queried	May 2024

Table C2: LLM Model Information