Automatic Large Language Models Creation of Interactive Learning Lessons

Jionghao Lin[1,2,3], Jiarui Rao[2], Sandy Yiyang Zhao[2], Yuting Wang[2], Ashish Gurung[2], Amanda Barany[4], Jaclyn Ocumpaugh[4], Ryan S. Baker[4], and Kenneth R. Koedinger[2]

[1] The University of Hong Kong, Pokfulam Rd, Hong Kong, China
jionghao@hku.hk
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Monash University, Wellington Rd, Clayton VIC 3800, Australia
[4] University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract. We explore the automatic generation of interactive, scenario-based lessons designed to train novice human tutors who teach middle school mathematics online. Employing prompt engineering through a Retrieval-Augmented Generation approach with GPT-4o, we developed a system capable of creating structured tutor training lessons. Our study generated lessons in English for three key topics—Encouraging Students' Independence, Encouraging Help-Seeking Behavior, and Turning on Cameras—using a task decomposition prompting strategy that breaks lesson generation into sub-tasks. The generated lessons were evaluated by two human evaluators, who provided both quantitative and qualitative evaluations using a comprehensive rubric informed by lesson design research. Results demonstrate that the task decomposition strategy led to higher-rated lessons compared to single-step generation. Human evaluators identified several strengths in the LLM-generated lessons, including well-structured content and time-saving potential, while also noting limitations such as generic feedback and a lack of clarity in some instructional sections. These findings underscore the promise of hybrid human–AI approaches in creating lessons for tutor training.

Keywords: Generative Artificial Intelligence, Large Language Models, Content Generation, Professional Development, Lesson Design

1 Introduction

One-on-one human tutoring is widely recognized as an effective strategy for enhancing student learning, with substantial evidence highlighting its positive impact on learning outcomes [18]. However, scaling human tutoring remains a significant challenge due to the scarcity of skilled tutors [18]. To address this issue, professional development programs have been developed to equip novice and nonprofessional tutors with the skills needed for effective instruction [18]. Online scenario-based training has emerged as a promising approach, simulating real-life tutoring situations to prepare novice tutors for their roles [22]. Central to

scenario-based tutor training program is well-constructed lessons that offer specific instructional strategies and practical guidance for adapting human tutoring to diverse learner needs. However, developing such lessons manually requires significant time and expertise, making the lesson design process costly and difficult to scale. These challenges underscore the urgent need for artificial intelligence (AI) technologies for automatic lesson generation, which can streamline lesson development and expand access to effective tutor training programs.

Lesson generation using AI techniques has been investigated in previous research to address the challenges of the time-consuming and labor-intensive process of creating educational materials. Early approaches employed rule-based systems to generate quiz questions, providing a foundational method but lacking adaptability and depth [24, 10]. Subsequent approaches introduced template-based methods and early AI models, which improved the variety and relevance of generated content [19]. However, these approaches still faced significant limitations, including inflexibility and a lack of contextual understanding [19].

Recently, large language models (LLMs) demonstrate exceptional capabilities in generating human-like text, leveraging pretraining on extensive datasets [20]. These capabilities enable LLMs to potentially address the shortcomings of earlier approaches by producing nuanced, contextually relevant, and adaptable content, making them particularly well-suited for educational purposes. However, generating comprehensive lessons for tutor training remains a complex task. A lesson typically consists of multiple interconnected segments, including well-defined learning objectives and quiz. Each segment serves a specific instructional purpose and must align with the overall pedagogical goals of the lesson. Given the complexity of lesson design, generating all components in a single step risks overwhelming the model, resulting in shallow or disjointed outputs [9].

To address this, we adopt a prompting strategy similar to [25], chaining multiple LLM steps together—where the output of one step serves as the input for the next—and has been shown to enhance the quality of task outcomes. Research highlights the benefits of such decomposition strategies for tasks requiring multi-step reasoning, as they help improve the quality of generated content [25, 9]. However, it remains unknown how extensively a task should be decomposed to ensure optimal performance, since overly complex sub-goals might still overwhelm the LLM [13]. In the context of lesson generation, this challenge becomes particularly relevant, as lessons consist of multiple interdependent components that must align both structurally and pedagogically. It remains unclear how the number of segments or the level of decomposition impacts the quality, coherence, and pedagogical rigor of the generated lessons.

The demand for high-quality lesson content in tutor training programs remains substantial; however, the development of content specifically tailored for tutor training has been relatively under-explored. Our study developed a lesson content generation system that leverages prompt engineering with LLMs. The system was designed to generate scenario-based lessons for training novice tutors who teach middle school mathematics in online environments. To evaluate the effectiveness of the generated lessons, we collaborated with human instructional

designers, who evaluated the lesson quality using a comprehensive rubric. This evaluation aimed to determine how effectively LLMs can generate lesson content that meets the pedagogical and practical needs of tutor training. Based on this investigation, our study proposed two **R**esearch **Q**uestions (**RQs**): **RQ1:** *How does the level of task decomposition influence the quality of tutor training lessons generated by LLMs?* **RQ2:** *How does the quality of lessons generated by LLMs compare to those created by expert human lesson designers?*

2 Related Work

2.1 Scenario-based Lesson Development for Tutor Training

Scenario-based lessons are a key component of effective tutor training, providing a realistic context for tutors to apply their skills [4]. These lessons simulate authentic tutoring interactions, helping tutors understand and navigate common challenges they might face. By working through scenarios that mimic real-world situations, tutors develop the flexibility needed to effectively support diverse learners. Additionally, lessons focusing on advocacy, empathy, and cultural competence are important for preparing tutors to work with diverse student populations [22]. These competencies enable tutors to deliver academic content and address students' emotional well-being and equity-related challenges, fostering supportive learning environments [4].

Lessons structured around "learning by doing" methodologies, such as the modified Predict-Observe-Explain (POE) model, have shown significant success in enhancing tutor preparedness [11]. In the scenario-based lessons, tutors first engage with a scenario by predicting the best response and then explaining their rationale. This approach fosters active engagement and critical reflection, helping tutors develop a deeper understanding of effective tutoring strategies [1]. Moreover, incorporating a combination of multiple-choice questions (MCQs) and open-ended questions within these lessons has proven effective for developing both basic and advanced tutoring skills [2]. MCQs help in assessing foundational knowledge in an efficient manner, while open-ended questions encourage deeper cognitive engagement, enabling tutors to articulate their understanding [8].

2.2 Lesson Generation

Lessons designed for tutor training are essential for equipping tutors with the skills needed to support diverse learners. However, manually crafting these lessons can be both time-consuming and labor-intensive, particularly when scaling tutor training programs to meet growing demand [18, 22]. Automated lesson generation approaches have emerged as a potential solution to these challenges, offering benefits such as time savings, scalability, and consistency in instructional content. Traditional methods of automated learning content generation, such as rule-based approaches or template-based approaches [6], often lack the flexibility and contextual depth. These limitations highlight the need for more advanced approaches to generating structured, pedagogically rich lessons.

The rise of LLMs has brought significant advancements to content generation in education. LLMs have demonstrated strong performance across a range of educational tasks, including generating feedback [17, 5], course learning topics [14] and questions [16]. For instance, GPT models have been used to provide feedback on open-ended student responses, producing actionable suggestions that support learning [5]. Despite the growing interest in LLMs for education, their use in generating structured, scenario-based lessons tailored to tutor training remains under-explored. Given the unique requirements of tutor training, the potential for LLMs to automate high-quality lesson generation at scale presents a significant opportunity.

3 Methods

3.1 Lesson Structure from a Tutor Training Platform

To generate the lessons, our study builds upon an existing tutor training program structured around the SMART framework demonstrated in [4, 22], which encompasses five key competencies critical for successful tutoring: **S**ocial-emotional Learning, **M**astery of Content, **A**dvocacy, **R**elationship-building, and **T**echnology. All lessons in our study were generated in English. The lessons are divided into five main sections, each designed to support specific aspects of tutor training:

- **Title Page:** This section provides a description of the lesson topic, including the key focus areas and the learning objectives the lesson aims to achieve.
- **Scenario I Section:** In this section, tutors are presented with the first tutoring scenario and are required to answer open-ended questions and multiple-choice questions (MCQ). These questions are designed to replicate real-world tutoring situations, asking tutors to respond as they would during an actual session and to justify their responses.
- **Instruction:** This section offers research-backed instruction for effective tutoring practices, helping tutors understand both desired and undesired tutoring practice. Each instruction section includes specific examples that help clarify best practices.
- **Scenario II Section:** This section follows a structure similar to the Scenario I section but presents a different scenario that requires tutors to apply their learned skills in a slightly altered context. This variation ensures that tutors understand the applicability of concepts beyond a single situation.
- **Conclusion:** The lesson concludes with a summary that reviews the key concepts covered throughout the lesson, reinforcing the understanding of tutoring strategies discussed. The conclusion section also includes a list of references, providing tutors with additional resources for deeper exploration

3.2 Human-Crafted Lesson

We invited two lesson designers to craft the lessons for tutor training and both lesson designers followed the structured lesson format outlined in Section 3.1. Each designer was tasked with creating lessons aligned with specific tutoring scenarios. The first lesson designer developed two lessons. The first lesson is

"Encouraging Students' Independence" which focuses on fostering student autonomy by training tutors to use polite, constructive language to guide learners without undermining their independence and the second lesson is *"Turning on Cameras"*, which aims to equip novice tutors with strategies to encourage camera use in online tutoring sessions, aiming to improve student engagement and active participation. The second lesson designer developed one lesson, i.e., *"Encouraging Help-Seeking Behavior"*, which aims to train tutors in identifying and addressing barriers that prevent students from seeking help, fostering a more supportive learning environment. All three lessons underwent rigorous review and multiple iterations, ensuring they were polished and ready for publication on the tutor training platform. The topics of these three human-crafted lessons were subsequently used to prompt the LLMs, following the prompt engineering approaches detailed in Section 3.3. The human-crafted lessons are publicly accessible through the OSF repository

3.3 Large Language Model-Generated Lessons

We aim to generate comprehensive tutor training lessons based on research articles related to effective tutoring practices. For generating more topic-specific content, Retrieval-Augmented Generation (RAG) has emerged as a powerful technique [12, 7]. RAG-based LLM content generation combines both retrieval and generation capabilities, allowing the model to first retrieve relevant documents and then utilize the relevant information to generate content. The strength of RAG lies in its ability to produce contextually accurate content by incorporating the generated output into factual data, which is particularly advantageous when generating educational content that requires specific knowledge from reliable sources [21]. As shown in Figure 1, we propose a two-step approach for lesson generation. Firstly, a lesson designer retrieves articles related to specific tutoring practices, providing the foundational knowledge on effective instructional strategies. Secondly, these retrieved articles are used to prompt the GPT-4o model (`gpt-4o-2024-05-13`). The GPT-4o model has been recognized for its balance of cost efficiency and strong performance in content generation tasks, making it a suitable choice for producing pedagogically sound lesson content.

An important aspect of generating tutor training lessons is determining the optimal prompting strategy for content generation. As discussed in Section 3.1, each lesson consists of five main sections. It is important to explore whether it is more effective to generate an entire lesson in one step or generate multiple segments of lesson section (e.g., generate each section independently). As shown in Figure 1, the lesson structure includes five sections, and each section can be generated using different prompting strategies. Additional details about our prompting strategies can be accessed through the OSF repository

- **One Segment:** The entire lesson (all five sections) is generated at once.
- **Two Segments:** The lesson is split into two main segments for generation. The content from S1, including the title page and the description of Scenario I, is incorporated as part of the prompt for generating S2.

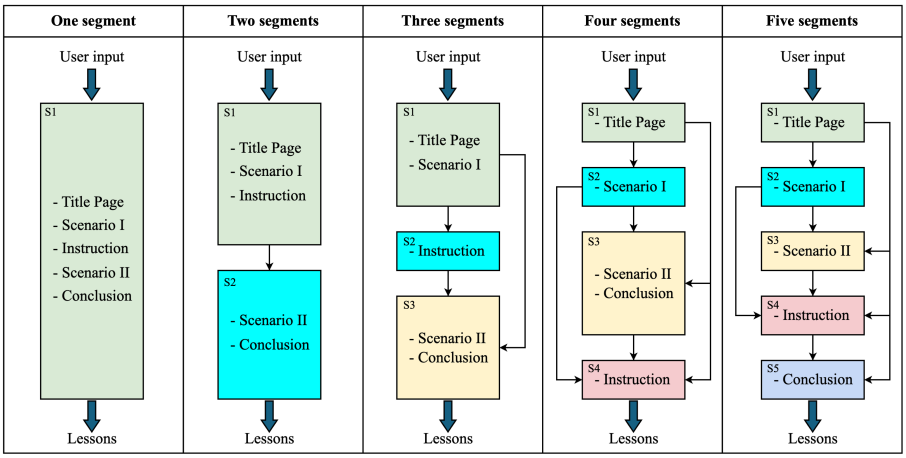


Fig. 1. Our approaches for prompting LLM to generate lessons for tutor training.

- **Three Segments:** The lesson was divided into three segments. The content from S1 is used to prompt the generation of the instructional content (S2), while both S1 and S2 are incorporated as inputs to generate Scenario II and the conclusion (S3).
- **Four Segments:** The lesson was split into four segments. Since the instruction section includes descriptions of research-recommended tutoring practices, generating it independently enhances clarity and depth. Scenario II and the Conclusion are generated together to streamline transitions.
- **Five Segments:** The most granular approach involves generating each of the five segments. This approach provided the highest level of detail and customization for each section, allowing the model to focus exclusively on the objectives of a single segment.

3.4 Lesson Evaluation

To evaluate the influence of different prompting strategies for lesson generation (RQ1), we hired two human coders to rate the content generated by the LLM. Both coders were experienced in tutoring middle school math subjects and had completed over ten tutor training lessons on the tutor training platform. The coders followed a detailed rating rubric, which was summarized from previous research [23] and included 17 distinct codes for evaluating lesson quality. Due to the page limit, the rating rubric is illustrated in Table A1 in Digital Appendix. To ensure rating consistency, both coders participated in multiple training sessions where they collaboratively rated a sample of LLM-generated lessons. These sessions continued until the coders achieved sufficient agreement on the application of the codes. After training, the two coders independently rated the LLM-generated lessons. The inter-rater reliability was assessed using Cohen's κ, which resulted in a score of 0.72. This score indicates substantial

agreement [15]. Any discrepancies were addressed through discussion involving both coders and a third reviewer, who also had experience tutoring middle school math and had completed over ten tutor training lessons on the platform. This process ensured that the final ratings accurately reflected the quality of the generated lessons. To answer RQ2, we collected feedback from two lesson designers using a structured written questionnaire. The designers were provided with a set of open-ended questions (see Table A2 in Digital Appendix). Each designer independently provided written responses to all questions.

4 Results

4.1 Analyze the quality of LLM-generated tutor training lessons

The results of the human rating scores (Table 1) demonstrate that the three-segment approach (**Three Seg.**) achieved the highest average rating (14.67) for lesson generation, while the one-segment approach (**One Seg.**) received the lowest average rating (10.67). These findings indicate that decomposing the lesson generation task into multiple subtasks can enhance the quality of the generated content. Interestingly, the five-segment approach (**Five Seg.**), which represents the highest level of task decomposition, performed slightly worse than the three-segment approach.

Table 1. Rating scores for LLM-generated tutor training lessons in different segments.

Lessons	**One Seg.**	**Two Seg.**	**Three Seg.**	**Four Seg.**	**Five Seg.**
Encouraging Students' Independence	12	11	13	13	13
Encouraging help-seeking behavior	8	11	15	15	12
Turning on Cameras	12	14	16	14	15
Average	10.67	12	14.67	14	13.33

Note: **Seg.** denotes **Segment**.

To gain a deeper understanding, we further analyzed the details of the rating scores for each section across the segment approaches. The scores in Table 2 reflect the final consensus ratings across the three generated lessons for each segmentation approach. For each criterion (e.g., *Specific* or *Clarity*), a score of 3/3 indicates that all three lessons met the standard for that code, while a score of 0/3 means that none did. In the **Title page** (see Table 2), across all segments, the code *Specific* achieved consistently high scores, indicating that the generated content effectively specified what the trainees would be able to do and under what conditions they would demonstrate the behavior. However, significant variability was observed in the code *Clarity*. In particular, the language used to define learning objectives often lacked clarity in certain segment configurations, such as the one-segment, two-segment, and five-segment approaches.

For **Scenario 1** and **Scenario 2**, the code *Problem Structure* consistently received high ratings across all segment prompts, highlighting the model's ability to replicate the structured format of tutor training lessons effectively. Additionally, both scenarios demonstrated strong alignment with learning objectives and

Table 2. Consensus ratings for each lesson section across different segmentation

Evaluation Codes	One Seg.	Two Seg.	Three Seg.	Four Seg.	Five Seg.
Title page					
Specific	3/3	3/3	3/3	3/3	3/3
Clarity	1/3	0/3	2/3	2/3	1/3
Scenario 1					
Problem Structure	3/3	3/3	3/3	3/3	3/3
Alignment with LO	2/3	3/3	3/3	2/3	2/3
Feedback	1/3	1/3	3/3	2/3	2/3
Plain Language	3/3	3/3	3/3	3/3	3/3
Instruction					
Alignment with LO	3/3	3/3	3/3	3/3	3/3
Clarity of Writing	0/3	0/3	1/3	2/3	2/3
Plain Language	1/3	2/3	3/3	3/3	3/3
Use of Examples	3/3	3/3	3/3	3/3	3/3
Pedagogy Grounded	0/3	1/3	3/3	1/3	1/3
Scenario 2					
Problem Structure	3/3	3/3	3/3	3/3	3/3
Alignment with LO	2/3	3/3	3/3	3/3	3/3
Feedback	1/3	2/3	2/3	2/3	2/3
Plain Language	3/3	3/3	3/3	3/3	3/3
Conclusion					
Alignment with LO	3/3	3/3	3/3	3/3	3/3
Authenticity of References	0/3	0/3	0/3	1/3	0/3

Note: **Seg.** denotes **Segment**. Each score (e.g., 3/3) represents the number of generated lessons (out of three) that received a positive consensus rating for the given criterion and segmentation.

employed clear, accessible language, as reflected in the high ratings for the codes *Alignment with Learning Objectives (LO)* and the *Plain Language* across all segment prompts. The provision of feedback (code *Feedback*) showed notable improvements in prompts using three or more segments compared to the one- or two-segment approaches. This improvement is likely attributed to the decomposition of content generation tasks, which enables the LLM to allocate more attention to generate feedback-related content

In the **Instruction** section, the LLM-generated lessons also demonstrated strong alignment with the learning objectives and provided practical examples. However, the code *Clarity of Writing* received lower scores for the one- and two-segment approaches. This limitation likely stems from the generation process, where the instruction section was integrated into a larger task that combined multiple components (see Figure 1). In contrast, the instruction section was generated independently in the three-, four-, and five-segment approaches, which allowed the LLM to focus more on delivering specialized concepts with detailed explanation derived from research-based articles. The code *Pedagogy Grounded* also showed variability across segment configurations, receiving lower scores across all approaches except the three-segment setup. Pedagogical grounding, which assesses the integration of theoretical principles into practical instructional guidance, was most effective in the three-segment approach. This improvement might be attributed to the sequential generation process, where

the instruction section was generated earlier in the three-segment strategy compared to the four- and five-segment approaches. Further investigation is required to better understand the underlying reasons for the lower scores in *Clarity of Writing* and *Pedagogy Grounded*, which will help refine the prompting strategies to achieve more consistent performance across all sections of the lesson.

Finally, the **Conclusion** section consistently underperformed in generating authentic references across all segment configurations, as reflected by the low scores for the code *Authenticity of References* in Table 2. Even with the optimal segment configuration (i.e., three segments) combined with the RAG approach—which integrated academic papers containing verified references—the model still produced references that did not exist, as illustrated in Table A3 in Digital Appendix. This finding highlights that while the RAG approach enhances the contextual relevance of generated content by incorporating retrieved documents, it does not inherently address the issue of citation authenticity.

4.2 Compare LLM-generated Lessons with Human-Crafted Lessons

To answer **RQ2**, we invited two lesson designers who were asked to identify strengths and weaknesses of these generated lessons compared to their manually crafted lessons, which encouraged critical analysis rather than subjective preference. Building on the findings from RQ1, we selected lessons generated using the three-segment prompt for comparison, as it generally demonstrated optimal performance across different lessons and codes as shown in Table 1 and 2. The reflections of lesson designers revealed several positive aspects of LLM-generated lessons, highlighting the system's utility in lesson design.

Strengths of LLM-generated Lesson Time-saving in creating lesson: One of the most significant advantages of LLM-facilitated lesson generation, as highlighted by both lesson designers, is the substantial time savings in drafting scenario sections. Crafting effective scenarios is a difficult process that requires maintaining a careful balance between similarity and distinctiveness across questions. As described in Section 3.1, Scenario I and Scenario II in a lesson must be comparable in difficulty and context to provide consistent learning opportunities, yet distinct enough to prevent students from perceiving them as identical. Lesson designer 1 elaborated on the challenge, noting, *"...The process of authoring and adjusting is incredibly time-consuming. The AI can quickly generate variations of scenarios, allowing educators to focus more on fine-tuning and less on repetitive drafting..."* Similarly, Lesson designer 2 emphasized the benefit of LLM assistance in creating multiple-choice questions within these scenarios, stating, *"...I think the LLM-generated lessons do save a lot of time from researching and creating the multiple choice questions."* Both designers agreed that this time-saving feature significantly reduces the cognitive and logistical load associated with scenario design. Additionally, designer 1 further commend *"...the instructional charts that's generated are mostly very helpful. ..."* as shown in Figure 2. The feedback from the two lesson designers suggests that the LLM-generated lessons can enhance the lesson creation workflow.

Strategy	Description	Good Example	Bad Example
Polite Request	Encourage initial engagement with polite requests.	"Could you please show me how you've tried to solve this problem so far?"	"Tell me what you've done on this problem."
Balanced Feedback	Provide polite feedback initially, then shift to direct feedback as needed.	"That's a good start, Sam. Let's see if we can fix this part together."	"That's wrong. You need to fix this part."
Assess Progress	Assess students' prior progress to tailor instructions.	"What steps have you taken so far on this problem?"	"Why haven't you made any progress on this problem yet?"

Fig. 2. LLM-generated examples of strategies for the lesson *Encouraging Students' Independence*. The chart highlights three strategies—Polite Request, Balanced Feedback, and Assess Progress—along with their descriptions, good examples of implementation, and corresponding bad examples to avoid.

Generating diverse and real-world scenarios: Lesson designer 1 noted, *"The AI can quickly generate variations of scenarios,"* and *"I think the scenarios are generally very-well generated. When given a strategy, GPT seems to be able to give scenarios where that strategy can be used, which I think is pretty helpful"* highlighting the model's capability to provide multiple scenarios tailored to different instructional strategies. For example, in the lesson focused on Encouraging Students' Independence, the LLM-generated content in both scenario sections:

- **Scenario I:** *"Tutor Alex is working with a student named Jamie, who often waits for direct answers rather than attempting to solve math problems independently. Jamie has shown some progress but still lacks confidence in tackling problems alone."*
- **Scenario II:** *"Tutor Taylor is working with a student named Sam, who frequently asks for hints and confirmations before making any progress on math problems. Sam has shown the ability to solve problems but often seeks reassurance, which hinders independent problem-solving."*

The examples above demonstrate how subtle variations in the scenarios can address distinct aspects of encouraging independence—one focusing on building confidence in hesitant students (**Scenario I**) and the other on managing students' reliance on reassurance (**Scenario II**).

Unidentified biased or offensive content: Ensuring unbiased and appropriate AI-facilitated learning content generation is essential for building trust among educators and learners. Lesson designer 1 highlighted this positive aspect, stating, *"I haven't seen any biased or offensive content in GPT-generated lessons."* Similarly, Lesson Designer 2 affirmed, *"I haven't seen this,... do not seem to include biased or offensive material."* The feedback from both lesson designers suggests that the LLM-generated lesson has avoided inappropriate content.

Facilitating Human-AI Collaboration: As acknowledged by both lesson designers, the AI-generated content provided a foundational draft that could be iteratively improved, significantly reducing the time and effort required to create lessons from scratch. Lesson designer 1 emphasized the value of this col-

laboration, noting, *"...The AI can quickly generate variations of scenarios, allowing educators to focus more on fine-tuning and less on repetitive drafting..."* and similarly, lesson designer 2 indicated, *"The quality of some LLMs generated lessons are good since they can be used after a few rounds of modification. The lessons I create also have been reviewing and polishing for several rounds..."* These observations underscore the collaborative potential of LLMs in lesson preparation and highlight a synergistic approach: AI efficiently handles repetitive and time-intensive tasks, while human expertise ensures the final output aligns with instructional best practices and meets the unique needs of learners.

Weaknesses of LLM-Generated Lessons While the positive aspects of LLM-facilitated lesson generation highlight its potential to streamline the process, reduce repetitive tasks, and generate realistic scenarios, it also has limitations. Lesson designers identified several areas where AI-generated content falls short of the standards achieved in human-crafted lessons.

Generic feedback content: Unlike handcrafted lessons, which often include detailed and targeted feedback for each possible answer, the AI-generated feedback content mostly focused on explaining the correct answer. This approach overlooked the importance of addressing why specific options were incorrect, leaving gaps in the learning experience. Lesson designer 1 noted, *"...Handcrafted questions often provide targeted feedback for each possible answer, explaining why a specific choice is correct or incorrect. In contrast, LLM-generated courses tend to focus only on explaining the correct answer, neglecting the nuanced reasoning behind other options. In the cases where an explanation for a wrong answer is provided, the explanation given is generally more one-size-fits-all style, which lacks that important factor of targeted feedback..."* Lesson designer 2 indicated *"The explanation for options 2, 3, 4 is not specific enough, since it doesn't explain exactly why we don't choose them."* and provided a specific example of LLM-generated feedback (see Table A4 in Digital Appendix). These observations highlight the necessity of incorporating more nuanced explanations for all answer options, both correct and incorrect. Lacking detailed feedback for incorrect answers diminishes the educational value of the lesson. Targeted feedback is critical for helping learners understand the rationale behind their choices and correct misconceptions. The generic explanations provided by the LLM fail to address this need, resulting in less effective learning experiences.

Clarity in the title page and instruction section: Both designers noted instances where the generated lessons inconsistently used terms like "learners", "teachers", "tutors", and "students". This inconsistency sometimes confused the intended novice tutors, as the lessons were specifically designed for training tutors on how to teach their students effectively. As shown in Table 3, the examples generated by one-segment prompt fail to maintain clarity about the intended novice tutors—confusing "learners" with "tutors"—and does not clearly address the objectives of the tutor training lesson. Such inconsistencies can hinder novice tutors' understanding of how the training content applies to their practice.

Additionally, the instruction section of the LLM-generated lessons exhibited similar issues of lack of clarity. Lesson designer 1 highlighted that *"...the gen-*

Table 3. Examples of learning objectives generated using three-segment and one-segment prompts

Three-segment generation	One-segment generation
"Explain how polite instructional strategies can improve student engagement and problem-solving in online math tutoring."	*"Describe how learners will interpret or explain the tutoring strategy."*
"Develop and implement a balanced approach to using polite and direct instructions to foster student independence."	*"Outline how learners will generate or implement the tutoring strategy."*

erated instructions are overly long and lack a logical backbone to support the reasoning behind the strategies..." This disorganization often made it difficult for users to follow the intended flow, as instructions were included excessive details and citations. Additionally, lesson designer 1 added *"...There are some cases where the length and citations used are too much, making it difficult to pick out the emphasis of the instruction..."* An example of this overuse of citations can be seen in Table A3 in Digital Appendix. Lesson designer 2 echoed similar concerns, emphasizing the structural and logical disconnect in the instruction sections. *"...The flow and logic can be more natural...."*. This lack of cohesive structure often led to instructions that felt fragmented and harder to comprehend. *"...I find it difficult to read and understand since the logic can be a bit disconnected..."*. An example of this issue can be found in the following excerpt from a generated lesson: "*...Wallace, Hand, and Prain (2004) highlighted the effectiveness of the Science Writing Heuristic (SWH) in improving conceptual understanding through student-generated questions...*" The lack of context or explanation about what the SWH framework entails makes it difficult for readers unfamiliar with the framework to understand its relevance. This omission can leave readers, particularly novice tutors, struggling to understand how such frameworks apply to their tutoring practices.

5 Discussion and Conclusion

Our study investigated the potential of using large language models (LLMs), specifically with the retrieval-augmented generation (RAG) approach, to automate the generation of tutor training lessons for middle-school math pedagogy. Through structured evaluations, we assessed the quality, strengths, and limitations of LLM-generated lessons. Our findings underscore the importance of task decomposition in lesson generation. By decomposing the task into manageable subtasks, the LLM could maintain logical coherence across lesson components, demonstrating consistency with prior research advocating task decomposition strategies [25, 9]. Conversely, the one-segment approach, which involved generating the entire lesson within a single prompt, resulted in lower-quality outputs. Interestingly, the five-segment approach, representing the highest level of decomposition, introduced challenges such as reduced clarity and weaker pedagogical grounding. These challenges highlight the need for further investigation to determine whether they stem from excessive decomposition, potentially disrupting the logical flow and consistency of the generated content, or from rating bias

or manual rating inconsistencies that may have influenced the findings. Moreover, our study offers practical insights for using LLMs in educational content creation, demonstrating their ability to generate diverse and realistic lesson scenarios, establish the ethical reliability of the generated lessons, and support their value in scaling inclusive educational resources.

Despite these advantages, our investigation identified key areas where LLM-generated content requires further improvement. Feedback provided for open-ended and multiple-choice questions frequently lacked detailed explanations necessary for addressing misconceptions related to incorrect responses. Furthermore, issues with clarity were evident in the lesson content including title page (e.g., unclear descriptions of learning objectives) and instructional sections (e.g., logical disconnects, and insufficient explanation of jargon). These shortcomings highlight the need for refined prompting strategies to improve precision and coherence in generated content. Another weakness of LLM-generated lessons is the authenticity of academic references, which were often unreliable. Human lesson designers still need to manually verify and supplement references if lessons require source citations. The strengths and weaknesses of LLM-generated lessons emphasize that while LLMs hold immense potential for automating lesson generation, their effectiveness is maximized within a human-AI collaborative workflow. Educators and instructional designers play a crucial role in reviewing and refining AI-generated content to ensure it aligns with pedagogical objectives and meets the nuanced needs of learners.

Future Works. Future research should focus on designing more targeted and specific technical approaches to address limitations identified in LLM-generated lessons. For instance, a multi-agent approach could assign specialized agents to generate, evaluate, and refine lesson content, thereby improving feedback quality through iterative review [3]. Moreover, integrating multiple modalities—such as pairing AI-generated text with relevant visuals or slides—into lessons could better support diverse learning preferences in tutor training [26]. Additionally, our study's lesson evaluation rubric was primarily informed by one resource [23]. While the rubric [23] is comprehensive, future research could benefit from comparing or integrating it with established rubrics in the field to further validate and strengthen lesson evaluation. Then, enabling direct collaboration between instructional designers and the LLM during the refinement process can provide opportunities for real-time adjustments. For example, instructional designers could request immediate changes to tone, clarity, or content specificity, iteratively improving lesson quality. This interactive feedback loop would increase the quality of generated content and streamline the lesson creation process. Lastly, our study primarily focused on evaluating LLM-generated lessons based on human evaluators' judgments, serving as a foundational step in identifying high-quality lessons for further exploration. Future research should conduct experiments to evaluate the relative effectiveness of LLM-generated lessons compared to human-crafted ones, providing insights into their impact on participants' learning outcomes. The results would help clarify the extent to which LLM-generated lessons achieve comparable or complementary results to those designed by human experts.

Acknowledgment

This work was supported by the Learning Engineering Virtual Institute[5], and a grant from the University Research Committee (Grant No. 2401102970) at The University of Hong Kong. The opinions, findings, and conclusions expressed in this paper are solely those of the authors.

Bibliography

[1] Aleven, V.A., Koedinger, K.R.: An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. Cognitive science **26**(2), 147–179 (2002)
[2] Butler, A.C.: Multiple-choice testing in education: Are the best practices for assessment also good for learning? Journal of Applied Research in Memory and Cognition **7**(3), 323–331 (2018)
[3] Cao, J., Zhao, C.Q., Chen, X., Wang, S., Schunn, C., Koedinger, K.R., Lin, J.: From first draft to final insight: A multi-agent approach for feedback generation. arXiv preprint arXiv:2505.04869 (2025)
[4] Chhabra, P., Chine, D., Adeniran, A., Gupta, S., Koedinger, K.: An evaluation of perceptions regarding mentor competencies for technology-based personalized learning. In: Society for Information Technology & Teacher Education International Conference. pp. 1812–1817. AACE, San Diego, CA, USA (2022)
[5] Dai, W., Tsai, Y.S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., Chen, G.: Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. CAEAI **7**, 100299 (2024)
[6] Diwan, C., Srinivasa, S., Suri, G., Agarwal, S., Ram, P.: Ai-based learning content generation and learning pathway augmentation to increase learner engagement. CAEAI **4**, 100110 (2023)
[7] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
[8] Gurung, A., Vanacore, K., Mcreynolds, A.A., Ostrow, K.S., Worden, E., Sales, A.C., Heffernan, N.T.: Multiple choice vs. fill-in problems: The trade-off between scalability and learning. In: Proceedings of the 14th Learning Analytics and Knowledge Conference. pp. 507–517 (2024)
[9] Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., Sabharwal, A.: Decomposed prompting: A modular approach for solving complex tasks. arXiv preprint arXiv:2210.02406 (2022)
[10] Koedinger, K.R., Corbett, A.T.: Cognitive tutors: technology bringing learning science to the classroom. The Cambridge Handbook of the Learning Sciences pp. 61–77 (2006)
[11] Koedinger, K.R., Kim, J., Jia, J.Z., McLaughlin, E.A., Bier, N.L.: Learning is not a spectator sport: Doing is better than watching for learning from a mooc. In: Proceedings of the Second (2015) ACM Conference on L@S. pp. 111–120. ACM, Vancouver, BC, Canada (2015)

[5] https://learning-engineering-virtual-institute.org/

[12] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
[13] Li, H., Leung, J., Shen, Z.: Towards goal-oriented large language model prompting: A survey. arXiv preprint arXiv:2401.14043 (2024)
[14] Moein, M., Molavi Hajiagha, M., Faraji, A., Tavakoli, M., Kismihók, G.: Beyond search engines: Can large language models improve curriculum development? In: European Conference on Technology Enhanced Learning. pp. 131–136. Springer (2024)
[15] Neuendorf, K.A.: The Content Analysis Guidebook. SAGE Publications (2017)
[16] Nguyen, H.A., Bhat, S., Moore, S., Bier, N., Stamper, J.: Towards generalized methods for automatic question generation in educational domains. In: European Conference on Technology Enhanced Learning. pp. 272–284. Springer (2022)
[17] Nguyen, H.A., Stec, H., Hou, X., Di, S., McLaren, B.M.: Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In: European Conference on Technology Enhanced Learning. pp. 278–293. Springer (2023)
[18] Nickow, A., Oreopoulos, P., Quan, V.: The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. Working Paper 27476, National Bureau of Economic Research (July 2020). https://doi.org/10.3386/w27476
[19] Paladines, J., Ramirez, J.: A systematic literature review of intelligent tutoring systems with dialogue in natural language. IEEE Access **8**, 164246–164267 (2020)
[20] Radford, A.: Improving language understanding by generative pre-training. OpenAI technical report (2018)
[21] Rao, J., Lin, J.: Ramo: Retrieval-augmented generation for enhancing moocs recommendations. arXiv preprint arXiv:2407.04925 (2024)
[22] Thomas, D., Yang, X., Gupta, S., Adeniran, A., Mclaughlin, E., Koedinger, K.: When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors. In: LAK23. pp. 250–261. ACM, Arlington, TX, USA (2023)
[23] Vai, M., Sosulski, K.: Essentials of online course design: A standards-based guide. Routledge (2011)
[24] Wiemer-Hastings, P., Graesser, A.C., Harter, D., Group, T.R.: The foundations and architecture of autotutor. In: ITS. pp. 334–343. Springer (1998)
[25] Wu, T., Terry, M., Cai, C.J.: Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In: Proceedings of the 2022 CHI. pp. 1–22 (2022)
[26] Zhao, C.Q., Cao, J., Chen, E., Koedinger, K.R., Lin, J.: Slideitright: Using ai to find relevant slides and provide feedback for open-ended questions. arXiv preprint arXiv:2505.04584 (2025)

A Appendix

Table A1. Rating rubric for evaluating the quality of LLM-generated lesson

Section	Rubric	Maximum Rating (pts.)
Title Page	1) Specific: The learning objectives should clearly state what the trainees will be able to do and the conditions under which they will demonstrate these behaviors (1 pt.) 2) Clarity: The wording used to define learning objectives is clear. (1 pt.)	2 pts
Scenario (Scenario I & II)	1) Problem Structure: Each scenario consists of exactly two open-ended problems, which are both predicted and explained, along with two multiple-choice questions, also predicted and explained. (1pt.) 2) Alignment with Learning Objectives: All problems are directly aligned with the learning objectives, ensuring that they thoroughly cover key concepts. (1pt.) 3) Feedback: The provided feedback is clear and directly addresses the problem at hand. The feedback content is reasonable and accurate, ensuring that learners/trainees can easily follow and understand the material. (1pt.) 4) Plain Language: Uses familiar and common words whenever possible. Avoids jargon, clichés, and idiomatic expressions. Special terms, abbreviations, and acronyms are clearly defined and easily accessible. (1pt.)	4 pts for each Scenario
Instruction	1) Alignment with Learning Objectives: Instruction is comprehensive and directly aligns with all stated learning objectives. (1pt.) 2) Clarity of Writing: The writing demonstrates clarity, conciseness, and logical flow. Sentences and paragraphs are brief yet impactful, with well-structured ideas presented directly and to the point. (1pt.) 3) Plain Language: Uses familiar and common words whenever possible. Avoids jargon, clichés, and idiomatic expressions. Special terms, abbreviations, and acronyms are clearly defined and easily accessible. (1pt.) 4) Use of Examples: Presentations include high-quality examples, case studies, and illustrations that effectively clarify complex concepts. (1pt.) 5) Pedagogy Grounded: Instruction bridges theory and practice, enhancing student engagement. (1pt.)	5 pts
Conclusion	1) Alignment with Learning Objectives: Conclusion is comprehensive and directly aligns with all stated learning objectives. (1pt.) 2) Authenticity of References: References must be valid, authentic, and verifiable. (1pt.)	2 pts
Total pts.		17 pts

Table A2. Structured questions for collecting feedback from lesson designers.

Themes	Questions
Strengths and weaknesses of LLM-generated lessons compared to manual crafted lessons	1. How do the lessons you created using LLMs compare to those you created manually?
	2. Can you provide specific examples where LLM-generated lessons were much better than lessons you designed by hand?
	3. Can you provide specific examples where LLM-generated lessons were much worse than lessons you designed by hand?
	4. What is the most challenging thing you have faced when building a lesson by hand? What role might a large language model (LLM) play in addressing that challenge?
	5. What specific quality issues have you encountered with lessons generated by large language models (LLMs)? Could you describe these issues in detail with examples?
	6. Have you encountered instances where GPT-generated lessons appeared well-written and accurate at first glance, but contained errors or inaccuracies upon closer inspection? If so, can you provide examples?
Quality of LLM-generated scenario-based lessons for tutor training	1. How effectively do you think the title page of GPT-generated lessons cover the learning objectives?
	2. How accurately does the scenario context description in GPT-generated lessons reflect real-world tutoring sessions?
	3. How easy is it for you to select the correct answer from the options provided in GPT-generated lessons? Do the correct answers make sense to you?
	4. How easy do you find it to understand the instructional content (i.e., research recommendations) part of GPT-generated lessons?
	5. Have you seen evidence of biased or offensive content in GPT-generated lessons?
Further improvement for LLM-generated lessons	1. What improvements or features would you suggest for enhancing the quality of LLM-generated lessons?
	2. How can collaboration between LLMs and human instructional designers be improved?
	3. Is there anything else you would like to share about your experiences with LLM-generated lessons?
	4. Do you have any additional suggestions for researchers or developers working on improving LLMs for educational purposes?

Table A3. References generated in the lesson *Encouraging Students' Independence* using the three-segment prompt.

LLM-generated research-based instruction (shown in the Instruction Section): *Research has shown that students who receive polite tutorial instructions tend to perform better than those who receive direct instructions* ***(Schneider et al., 2022)****. This is particularly true for students with lower prior knowledge, as polite expressions can create a supportive learning environment that encourages engagement and persistence* ***(Mikheeva et al., 2022)****. However, excessive politeness can hinder instructional efficiency and may not be effective for students with higher prior knowledge, who may benefit more from direct instructions* ***(Brummernhenrich & Jucks, 2022)****. The effectiveness of polite instruction is also influenced by the students prior progress. Tutors should assess students progress at the beginning of the session to tailor their instructional strategies accordingly* ***(Yang & Li, 2018)****. For example, students who have made some progress may need less polite scaffolding and more direct feedback to push them further. Finally, an effective tutoring session often involves a dynamic shift from polite to direct instructions as the session progresses* ***(Lin et al., 2022)****. This gradual shift helps maintain a balance between providing support and encouraging independence, which is essential for developing students problem-solving skills*

In-text Citations (Instruction Section)	**LLM-generated References (Conclusion Section)**	**Hallucination Issues**
(Schneider et al., 2022)	Schneider, B., et al. (2022). Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues. LAK22.	Incorrect authors
(Mikheeva et al., 2022)	Mikheeva, A., et al. (2022). The Role of Politeness in Instructional Communication. Journal of Educational Psychology.	Incorrect year, title and venue
(Brummernhenrich & Jucks, 2022)	Brummernhenrich, B., & Jucks, R. (2022). The Impact of Politeness on Instructional Efficiency. Educational Research Review.	Incorrect year and title
(Yang & Li, 2018)	Yang, F., & Li, F. W. B. (2018). Study on Student Performance Estimation and Progress Analysis. Computers & Education.	Incomplete title
(Lin et al., 2022)	Lin, J., et al. (2022). Dynamics of Politeness in Effective Tutoring Sessions. Journal of Learning Analytics.	Incorrect title and venue

Table A4. Example of using structured question stems.

Which of the following principles best supports the strategy of using structured question stems?

A. Structured question stems guide students thinking and help them focus on specific aspects of the topic. (**Correct**)

B. Allowing students to explore topics freely without guidance encourages creativity. (**Incorrect**)

C. Memorizing questions ensures students can recall information during tests. (**Incorrect**)

D. Revisiting questions later helps students develop independent learning skills. (**Incorrect**)

LLM-generated Feedback: The correct answer is (1) because structured question stems provide the necessary scaffolding to guide students thinking, making their questions more focused and meaningful. Options (2), (3), and (4) do not align with the research recommendations.