Time and Trade-offs: A Preliminary Study of LLM Latency in Math Automatic Question Generation

Taisei Yamauchi[1,*], Brendan Flanagan[2] and Hiroaki Ogata[1]

[1]Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
[2]Graduate School of Information Science and Engineering, Ritsumeikan University, 2-150 Iwakura-cho, Ibaraki, Osaka, 567-8570, Japan

Abstract

This study investigates the impact of large language model generation latency on pedagogical quality and user perception within an LLM-driven mathematical question-generation system called PRIME. Testing with junior high school students revealed that the system's average generation time for a question and a solution was 57 seconds, significantly exceeding standard wait-time thresholds. Crucially, correlation analyses showed that longer latencies significantly correlated with lower expert-rated validity, answerability, and explainability, alongside decreased student comprehension and increased perceived difficulty. To address these structural challenges, architectural improvements for LLM learning support systems were proposed. These include decoupled delivery, which is to present questions immediately while computing answers in the background, dynamic model routing based on task complexity, and transforming inevitable wait times into active learning opportunities through brief, retrieval-based interim tasks to maintain cognitive engagement.

Keywords

Latency, automatic question generation, math learning, large language model, classroom deployment challenges

1. Introduction and Related Work

System response time has long been recognized as a critical factor influencing user experience and cognitive flow [1]. With the widespread adoption of large language models (LLMs), the impact of response latency has become increasingly complex, as recent study indicates that user perception and the quality of human-LLM interaction vary significantly depending on both the task type and the duration of the latency [2].

While users generally prefer rapid responses [2], generating high-quality, complex educational content, such as step-by-step mathematical scaffolding, requires extensive computational reasoning from the system [3]. However, recent surveys on efficient reasoning highlight that LLMs can sometimes fall into "overthinking," executing excessive computations that unnecessarily increase generation time without proportional gains in output quality [4]. Conversely, the concept of "positive friction" suggests that immediate responses are not always optimal; introducing intentional, moderate delays or cognitive hurdles can sometimes encourage deeper processing and improve the overall human-AI interaction [5].

Despite these broader investigations into LLM latency and interaction, a significant gap remains in specific educational applications. To date, no existing research has investigated the relationship between generation time, generation accuracy, and student perception in the context of personalized question generation based on mathematical scaffolding. Understanding how generation time alters the complexity and pedagogical effectiveness of the content, as well as how these changes are perceived by both experts and students, is essential for deploying effective real-time LLM systems, especially into the educational settings.

HAI-Agency: Workshop on Orchestrating Human and AI Agency for Proactive and Reflective Learning, co-located with the 27th International Conference on Artificial Intelligence in Education (AIED 2026), Seoul, Korea, June 27 - July 3, 2026.
*Corresponding author.
yamauchi.taisei.28w@st.kyoto-u.ac.jp (T. Yamauchi); flanagan@fc.ritsumei.ac.jp (B. Flanagan); hiroaki.ogata@gmail.com (H. Ogata)
https://sites.google.com/view/imachauy (T. Yamauchi); https://flanaganacademic.wordpress.com (B. Flanagan); https://sites.google.com/site/hiroakiogata (H. Ogata)
0000-0003-3688-2896 (T. Yamauchi); 0000-0001-7644-997X (B. Flanagan); 0000-0001-5216-1576 (H. Ogata)

Therefore, to address this critical gap and optimize LLM-assisted scaffolding systems, this study poses the following research questions:

RQ1: What is the time required to provide the question with the scaffolded question-generation system?

RQ2: How does the generation time of LLM-based educational materials impact the trade-off between item complexity and pedagogical effectiveness from both student and expert perspectives?

2. Experiment Framework

The experimental environment for this study was built upon a comprehensive digital learning ecosystem. Students accessed the target mathematical materials through BookRoll, an interactive online e-book platform [6].

The overall architecture of the system bridges these front-end interfaces with the educational infrastructure of the back-end. Technical integration was facilitated by FastAPI, which dynamically retrieved user profiles from the school's learning management system and synchronized them with the BookRoll environment. Within this established infrastructure, the PRIME system was deployed to handle the dynamic generation of computationally scaffolded review questions specifically. To ensure robust and comprehensive data analysis, all student interaction logs and system behaviors were securely archived in a MongoDB database.

2.1. PRIME

PRIME is a learning framework that leverages LLMs to dynamically synthesize review questions. The system's core mechanism involves decomposing a source mathematical question into its constituent "knowledge points" (the fundamental concepts or sequential steps required for a solution). Upon selecting a question, students self-assess their mastery of these points. Based on scaffolding ideas [7], PRIME then processes this input to generate scaffold questions that target specific knowledge gaps. For students who demonstrate mastery of all points, based on the idea of expertise reversal effect [8], the system generates complex, applied questions. Beyond question synthesis, PRIME provides real-time answer generation and collects multi-dimensional feedback, including perceived difficulty, linguistic fluency, and relevance, storing all interactions in a longitudinal learning history.

2.1.1. Classification of Learner Comprehension State and Determination of Generation Strategy

In this phase, the system decomposes a source mathematical question into its constituent knowledge points. Based on the learner's self-assessment of their mastery of these points, the system generates a binary sequence, modeling each point as either mastered or misunderstood. The optimal pedagogical intervention strategy is then determined based on this pattern. The content of the instructions given to the LLM is dynamically adjusted according to the targeted gaps. For instance, an all-mastered sequence triggers the generation of an applied, complex question, while continuous misunderstandings trigger scaffolded questions for fundamental review. This mechanism establishes the foundation for prompts tailored to the specific cognitive roadblocks of each learner.

2.1.2. Question Generation via Dynamic Prompt Construction

Based on the intervention strategy determined in the previous step, the system dynamically constructs a prompt for the LLM to execute the generation of a question. To ensure both generation quality and system stability, strict constraints are embedded within the prompt. Specifically, alongside inserting the context of the original question and the required knowledge points as parameters, the system rigidly defines the output format. This includes mandating the use of the MathJax format for mathematical

expressions and specifying output rules for certain characters. By submitting this structured prompt to the LLM, a new mathematical question focusing on the learner's weaknesses is generated.

2.1.3. Automated Verification and Delivery Assurance via Self-Solving Mechanism

To guarantee the mathematical validity and solvability of the generated question, the system executes a verification process wherein the LLM is prompted to solve the very question it just created. In this phase, a dedicated prompt is reconstructed to compel the LLM to output the solution process and the final answer as distinctly separated sections. The system evaluates whether the output adheres to the required syntactic format; if the requirements are not met, the system requests regeneration up to a strict limit of two iterations. Crucially, a generated question is presented to the learner after the self-verification, and the system has completely finalized its corresponding solution and explanation. This strict prerequisite eliminates the risk of presenting incomplete or unsolvable tasks, thus ensuring the robustness and educational quality of the generated question set.

3. Method

3.1. Data Collection

The study involved 120 ninth-grade students (aged 14–15) from three classes, all instructed by the same mathematics teacher. The participants were high-performing students compared to the country's average. This study was approved by the Institutional Review Board and conducted with the consent of legal guardians.

The experiment took place between July 11 and August 25, 2025. Regarding the learning materials, experts (junior high school math teachers and Master's students in science) selected four specific questions to serve as a target for the PRIME system. These questions were considered to provide a variety of difficulty levels (basic or advanced) and curriculum units (algebra or geometry). No teacher intervention was provided during the period to ensure that learning outcomes were derived solely from interactions with the system.

3.2. Question Review with PRIME

Students utilized the PRIME system to review the target questions. The review process followed sequential steps: selecting the target question, self-assessing their progress to determine a pre-review score, generating a review question in real-time, solving the generated question, checking the model answer, and recording a reflection. A total of 208 review processes were recorded during the experiment.

Internally, the system recorded the question generation time and the answer generation time separately in the log data. The OpenAI `o4-mini` API was employed to generate the review questions and their corresponding answers.

3.3. Evaluations

To investigate the relationship between question generation time and evaluations from both students and experts, a correlation analysis was conducted using the following metrics. Table 1 displays the specific question items for the evaluation.

- Student evaluation: In the reflection step, students answered five questions regarding the review item: 'Understanding', 'Rating', 'Difficulty', 'Fluency', and 'Relevance'. 'Understanding' and 'Fluency' were evaluated on a 3-point scale, while 'Rating', 'Difficulty', and 'Relevance' were evaluated on a 5-point scale.
- Expert evaluation: Two experts (a humanities undergraduate and a science graduate with high mathematical proficiency) assessed the quality of the AI-generated questions. The evaluation

rubric consisted of five categories: 'Grammar', 'Clarity', 'Validity', 'Answerability', and 'Explainability'. Each category was measured by a pair of items (10 items total) on a 3-point scale. Negatively worded items were reverse-coded.
- Analytical metrics: The analysis included the following indicators 'Len-quiz', 'Len-answer', and 'Pre-review score'.

To analyze the relationship between the continuous generation latency and the ordinal indicators, we calculated Kendall's rank correlation coefficient (τ_b), which robustly handles the frequent occurrence of tied ranks in Likert-scale data [9]. Furthermore, to evaluate the robustness of these estimates, 95% confidence intervals were computed using Fisher's Z-transformation adapted for Kendall's τ_b [10].

Table 1
Evaluation Metrics.

Subject	Category	Evaluation Content
Student	Understanding	To what extent was the question statement comprehensible to you?
	Rating	How effective was this task in supporting your learning review?
	Difficulty	How would you rate the cognitive challenge of this question?
	Fluency	Did the question text feel smooth and linguistically natural?
	Relevance	How well did this question align with the core concepts of the original question?
Expert	Grammar	The text contains grammatical or syntactical errors in Japanese.
		The sentence is appropriately structured as a mathematical interrogative.
	Clarity	The question description is overly complex and could be simplified.
		The instructional intent and the specific goal of the question are clear.
	Validity	There are logical inconsistencies or contradictions within the question.
		The provided model answer is factually and mathematically accurate.
	Answerability	The information provided is either insufficient or redundant for solving the task.
		The question leads to a unique and well-defined mathematical solution.
	Explainability	The vocabulary or expressions exceed the standard middle school curriculum.
		The overall text is tailored to the reading level of average middle schoolers.
Analytical	Len-quiz	Length of the generated question text
	Len-answer	Length of the generated answer text
	Pre-review score	The ratio of selected checks among the provided answer points

4. Result

4.1. The Time Required to Provide the Question with the Scaffolded Question-generation System

Table 2 presents the descriptive statistics for the generation times of the AI-generated questions and answers, and Figure 1 illustrates these durations using a box plot. Overall, the total generation process (TOTAL) took a mean time of 56.94 seconds, with a median of 50.06 seconds. The creation of questions (CREATE-Q) required an average of 22.16 seconds, whereas the creation of answers (CREATE-A) took a mean of 34.78 seconds.

Table 2
Creation Time of Question and Answers.

	mean	std	min	25%	50%	75%	max
CREATE-Q	22.16	12.68	5.40	14.34	19.02	24.53	89.29
CREATE-A	34.78	27.11	7.07	17.15	28.06	44.32	241.24
TOTAL	56.94	33.85	15.71	34.19	50.06	68.96	288.16

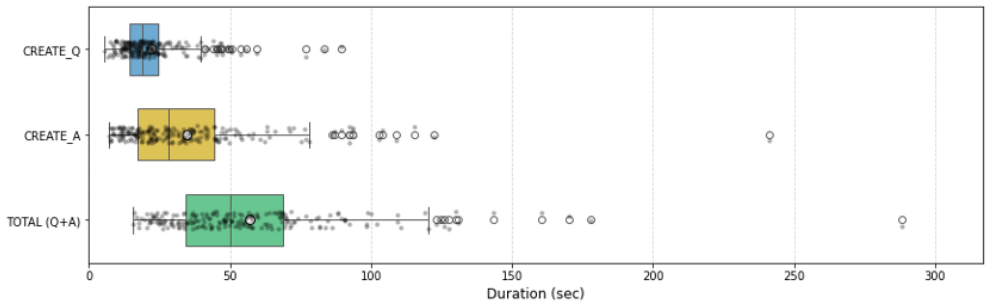

Figure 1: Box Plot of Creation Time of Question and Answers.

4.2. Correlation Between Creation Time and Student / Expert Evaluation of the Generated Questions

Table 3 describes the correlations between TOTAL creation time and metrics evaluated by students, by experts, or automatically. As the table shows, the indicators showed statistically significant correlations ($p < 0.001$) with this duration. Longer generation times were significantly associated with `Len-quiz`, `Len-answer`, and `Pre-review score`. Regarding student perceptions, an increase in generation time significantly decreased scores for `Understanding` and `Fluency`, while significantly increasing perceived `Difficulty`. Furthermore, expert evaluations revealed that longer generation times correlated with significantly lower scores for `Validity`, `Answerability`, and `Explainability`.

Table 3
Correlations between Generation Time and Various Indicators.

Indicator	Correlation coefficient (τ_b)	95% Confidence Interval
Understanding	-0.49***	[-0.57, -0.40]
Rating	-0.13*	[-0.22, -0.03]
Difficulty	0.38***	[0.29, 0.46]
Fluency	-0.29***	[-0.38, -0.20]
Relevance	-0.08	[-0.17, 0.02]
Grammar	-0.16**	[-0.25, -0.07]
Clarity	-0.15**	[-0.23, -0.06]
Validity	-0.27***	[-0.36, -0.19]
Answerability	-0.32***	[-0.40, -0.24]
Explainability	-0.18***	[-0.26, -0.09]
Len-quiz	0.43***	[0.36, 0.50]
Len-answer	0.44***	[0.37, 0.51]
Pre-review score	0.36***	[0.28, 0.43]

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5. Discussion and Conclusion

The average total generation time observed in this study (56.94 seconds) significantly exceeds the acceptable latency thresholds for learning support systems reported in prior literature. [2] characterized a 20-second latency in interactive systems as slow, while other frameworks suggest that feedback delivery within approximately 15 seconds is optimal for maintaining learner engagement [1]. These extended durations in the PRIME system likely stem from the inherent complexity of generating mathematically accurate content. While the results revealed a negative correlation where longer generation times were significantly associated with lower expert-rated validity and answerability, alongside higher perceived difficulty by students, it is crucial to recognize that this relationship is likely

confounded by task complexity. This suggests that when the LLM struggles with specific complex tasks, the excessive computational load simultaneously induces temporal delays and increases the risk of logical hallucinations or diminished pedagogical quality [4], rather than latency directly causing performance degradation.

To address these challenges, structural improvements to system architectures are necessary. First, systems can adopt a decoupled delivery mechanism. By presenting the generated question immediately while rendering answers in the background, the system could mitigate perceived latency. Furthermore, implementing dynamic model selection, which is not only use `o4-mini` but also stronger reasoning models like `gpt-5`, can help balance speed and reasoning capabilities, deploying higher-accuracy models only when task complexity demands it [11]. Additionally, unavoidable wait times can be pedagogically leveraged into active learning opportunities utilizing the "forward testing effect" [12]. By presenting brief retrieval-based interim tasks, systems can transform waiting periods into active engagement opportunities, alleviating the psychological burden while proactively optimizing the student's cognitive state.

This research suggests that latency in LLM-driven learning support systems might serve as a potential diagnostic indicator of pedagogical quality. It highlights the potential of integrating decoupled content delivery with dynamic model routing to match task complexity with appropriate model capabilities. Ultimately, the study points toward the necessity of transforming technical delays into active learning opportunities to improve instructional efficacy and student engagement.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3 in order to: Drafting content, Generate literature review, Plagiarism detection, Citation management, Grammar and spelling check. After using the tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. B. Miller, Response time in man-computer conversational transactions, in: Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I, AFIPS '68 (Fall, part I), ACM, 1968, pp. 267–277. doi:10.1145/1476589.1476628.

[2] F. F.-Y. Tan, M. A. Messerschmidt, W. Yin, O. Nov, The impact of response latency and task type on human-llm interaction and perception, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), ACM, 2026. doi:10.1145/3772318.3790716.

[3] F. Yu, H. Zhang, P. Tiwari, B. Wang, Natural language reasoning, a survey, ACM Comput. Surv. 56 (2024). doi:10.1145/3664194.

[4] Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, S. Zhong, N. Zou, H. Chen, X. Hu, Stop overthinking: A survey on efficient reasoning for large language models, arXiv preprint arXiv:2503.16419v4 (2025). URL: https://arxiv.org/abs/2503.16419.

[5] Z. Chen, R. Schmidt, Exploring a behavioral model of "positive friction" in human-ai interaction, in: A. Marcus, et al. (Eds.), HCI International 2024 – Late Breaking Papers, volume 14713 of *Lecture Notes in Computer Science*, Springer Nature Switzerland AG, 2024, pp. 3–22. doi:10.1007/978-3-031-61353-1_1.

[6] B. Flanagan, H. Ogata, Learning analytics platform in higher education in japan, Knowledge Management & E-Learning: An International Journal 10 (2018) 469–484. doi:10.34105/j.kmel.2018.10.028.

[7] D. Wood, J. S. Bruner, G. Ross, The role of tutoring in problem solving, Journal of Child Psychology and Psychiatry 17 (1976) 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x.

[8] S. Kalyuga, P. Ayres, P. Chandler, J. Sweller, The expertise reversal effect, Educational Psychologist 38 (2003) 23–31. doi:10.1207/S15326985EP3801_4.

[9] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93. doi:10.1093/biomet/30.1-2.81.

[10] E. C. Fieller, H. O. Hartley, E. S. Pearson, Tests for rank correlation coefficients. i, Biometrika 44 (1957) 470–481. doi:10.1093/biomet/44.3-4.470.

[11] D. B. Piskala, V. Raajaa, S. Mishra, B. Bozza, Dynamic llm routing and selection based on user preferences: Balancing performance, cost, and ethics, International Journal of Computer Applications 186 (2024) 1–7. doi:10.5120/ijca2024924108.

[12] C. Yang, D. R. Shanks, The forward testing effect: Interim testing enhances inductive learning., Journal of Experimental Psychology: Learning, Memory, and Cognition 44 (2018) 485. doi:10.1037/xlm0000449.