Toward Supporting Students' Metacognitive Questioning in GenAI-Assisted Learning

Yeonsun Yang
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
diddustjs98@dgist.ac.kr

Anastasia Arkhipenkova
Humanities, Arts, and Social Sciences, Yonsei University
Incheon, Republic of Korea
anastasia.rkh@yonsei.ac.kr

Jean Y. Song
Information and Interaction Design, Yonsei University
Incheon, Republic of Korea
jeansong@yonsei.ac.kr

Abstract

In the age of AI, asking a good question is becoming more important than figuring out the answer. In response, we collected and analyzed 1,150 questions that 92 university students asked to a GenAI tool while studying assigned STEM topics, using academic competence as an analytic lens. We categorized students' questions by content and investigated how these question types varied with academic competence. Our results show that metacognitive inquiries were more prevalent among high-performing students, suggesting that they engaged in more critical evaluation of AI-generated outputs and more proactive monitoring of their own understanding. In contrast, low-performing students tended to show greater overconfidence while reporting lower perceived mastery of their learning. Based on these findings, we propose design implications for future AI educational systems, including supporting learners' metacognitive regulation and preserving question generation as a learner-driven process. Additionally, we advocate for competence-sensitive scaffolding that adapts to students' questioning patterns without overriding their autonomy.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Metacognitive Questioning, Academic Competence, GenAI, STEM Education, University Students

ACM Reference Format:

Yeonsun Yang, Anastasia Arkhipenkova, and Jean Y. Song. 2026. Toward Supporting Students' Metacognitive Questioning in GenAI-Assisted Learning. In *Proceedings of CHI'26 Workshop on Understanding and Engaging Critical Resistance to AI in Education (CHI '26 Workshop).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Despite the productivity boost from generative AI (GenAI), recent studies report uneven benefits across users' task competence, which are reflected in distinct patterns of what and how they ask questions to AI [10, 14, 16, 19, 25]. For example, it is reported that GPT-4 improved high performers' entrepreneurial gains by about 20% but hindered low performers by roughly 10%, a disparity driven by the specificity and context of their questioning [19]. In programming education settings, higher-competency learners tended to ask questions that aligned with their existing knowledge, enabling more effective scaffolding from AI code generators [10]. In turn, the need to ask the good questions has catalyzed research on advancing AI literacy and cultivating prompt engineering skills [2, 14, 22].

Questioning is an integral component of scientific inquiry [27] and key metacognitive process [23] that goes beyond prompt engineering. In particular, self-questioning has been shown to enhance comprehension and academic performance [12]. Prior work has long emphasized the value of students' questions, which reveal their learning goals, recognized knowledge gaps, current knowledge states, the quality of thinking and reasoning, and efforts to extend and integrate new information with prior knowledge [1, 8, 15, 26]. In other words, the questions students pose to their learning partners (peers, teachers, and tools like search engines) can serve as a touchstone of their internal learning strategies and cognitive processes, thereby offering the potential to explain the divergent learning outcomes and experiences across individuals.

With the widespread adoption of GenAI tools in classrooms and self-directed learning, recent research in HAI and education has prompted efforts to examine the types and patterns of student-generated questions and the cognitive processes underlying them [6, 9, 21]. Yet, little is known about how students' questioning pattern differs by their academic competence. In this paper, we contribute to this discussion by investigating how university students use GenAI to support their STEM learning. We analyzed 1,150 conversation logs based on students' GPA and test scores, which were collected in our previous study [24]. Our findings suggest that good questions for stronger academic benefits are highly associated with *metacognitive regulation*—the critical evaluation of AI-generated content and proactive monitoring of one's knowledge states. Based on our results, we offer design recommendations for educators and AI developers to better support students' metacognitive questioning in GenAI-assisted learning.

2 Research Context and Approach

In the original study [24], we collected 1,150 conversation logs from 92 students at a university in South Korea (average age=21, 46 males and 46 females), and asked them to use ChatGPT-4o [18] to study selected STEM subjects. The students were instructed to study a randomly assigned topic (The Solar System, Sampling, or Database Management System) for 40 minutes, guided by three predefined learning objectives. To mitigate potential confounding effects, we

CHI '26 Workshop, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Our taxonomy for classifying the types of students' questions with ChatGPT-4o, capturing how they ask, request, state, and critique AI-generated outputs during self-directed STEM learning.

Category	Question Type	Explanation	Example
Information Inquiry	Concepts	Asking an explanation of a concept.	"*What is the Chi-squared distribution?*"(P3)
	Methods	Asking how processes, formulas, or principles work.	"*How does a planet's mass shape the way it evolves?*"(P38)
	Reasons	Asking about causes or underlying principles.	"*Why do planets orbit in elliptical paths?*(P26)"
Representation Inquiry	Elaboration	Requesting a more detailed or in-depth explanation of the current concept.	"*Could you explain statistical sampling in more detail?*"(P7)
	Rephrasing	Requesting the previous explanation to be rephrased.	"*Could you explain Kepler's third law again?*"(P17)
	Simplification	Requesting a simpler or more beginner-friendly explanation.	"*Explain RDBMS in a way that even a 3-year-old can understand.*"(P62)
	Material Generation	Requesting additional materials (*e.g.*, tables, diagrams, plots, or examples).	"*Explain Kepler's second law with a diagram.*"(P42)
	Summarization	Requesting a summary of previously given information.	"*Please summarize in one sentence.*"(P89)
Metacognitive Inquiry	Self-Explanation	Stating one's knowledge, opinion, or learning status.	"*I forgot the definition of a star, but I know it's a big object that makes its own light.*"(P39)
	Confirmation	Asking to confirm one's understanding.	"*So, is MySQL a type of RDBMS?*"(P87)
	Critique	Raising concerns or objections to a given response.	"*This diagram looks a bit odd—why is the planet placed at the center of the ellipse?*"(P39)
	Practice Problem Generation	Requesting the creation of practice problems to solve.	"*Based on the questions I've asked so far, please create nine test questions.*"(P50)
	Extending	Requesting further information connected to the current concept or prior knowledge.	"*Are there any other basic concepts I should know for understanding sampling?*"(P8)
	Learning Objectives	Restating or asking about learning objectives of the session.	"*I want to know [assigned learning objectives].*"(P39)
Miscellaneous	Social Expressions	Expressing greetings/goodbyes, gratitude, or other non-task-related social talk.	"*Thank you.*"(P71)
	Translation	Requesting a translation of the response into another language.	"*Say it again in Korean.*"(P2)

screened out students with prior knowledge of the topics. After the 40-minute study session, the students completed a closed-book test consisting of nine multiple-choice questions (scored out of 9 points) and predicted their own test scores. They also reported their perceived mastery of the topic using a 7-point Likert scale. Two weeks later, students took a retention test consisting of the same questions as the closed-book test, with randomized question and option order. We note that analyses of these extended data were not included in the original paper.

2.1 Derived Metrics

2.1.1 Academic Competence. To ensure a comprehensive indicator of academic competence, we z-normalized GPA, closed-book, and retention test scores and integrated them into a single composite score using Principal Component Analysis (PCA). Bartlett's test confirmed that the data were suitable for factor extraction. Parallel analysis supported a one-factor solution, with the first principal component (PC1) accounting for 48.9% of the total variance. Therefore, we used PC1 as the academic competence score, applying weights derived from the component loadings $W = [0.345, 0.684, 0.643]$ as follows:

$$\textit{Academic Competence} = W_1 Z_{\text{GPA}} + W_2 Z_{\text{Closed-book}} + W_3 Z_{\text{Retention}}$$

2.1.2 Metacognitive Calibration. Calibration represents the degree to which a student's perception of their performance matches with their actual performance [17, 20]. We calculated the calibration score as the confidence–competence gap [5], defined as the difference between the predicted and actual closed-book test scores (ranging from -9 to 9). Lower calibration scores indicate more accurate metacognitive monitoring.

2.2 Analysis

We conducted an open coding [3, 11] on students' questions asked to ChatGPT-4o during the study. Two of the authors jointly conducted all stages of qualitative coding. Following prior frameworks [6, 7, 9, 21], the authors initially classified the question data into similar codes first and defined names for the codes. The authors then iteratively reviewed the refined codes, merging similar codes and organizing them into higher-level categories. We carefully went through four rounds of coding to resolve all remaining discrepancies, excluding 59 questions that could not be reconciled and were omitted from the final analysis. In total, the 1,150 student-generated questions were classified into 16 codes and 4 categories.

Next, we examined how the frequency of each question type differed by academic competence scores using mixed-effects models. Additionally, we focused on high-performing (top 25% of participants) and low-performing students (bottom 25% of participants) and compared their questioning patterns and distribution. Finally, we compared their metacognitive calibration using independent samples t-tests and perceived mastery using Mann–Whitney U test.

3 Findings

This section presents the identified question types and the comparative results on questioning patterns and learning experiences.

3.1 More Metacognitive Questions were Asked Among High-Performing Students

In our taxonomy (Table 1), we catecorize metacognitive questions as stemming from students' metacognitive regulation, including queries that check and reflect on their understanding (self-explanation,

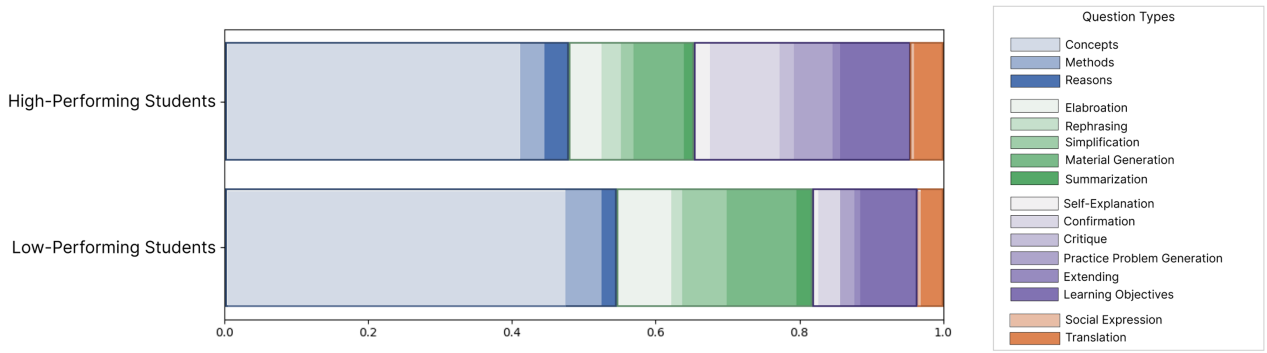

Figure 1: Distribution of question types asked by high- and low-performing students.

confirmation, practice problem generation), critically evaluate information (critique), and articulate or refine learning goals (learning objectives, extending). Overall, qualitative coding results showed that students engage with GenAI for STEM learning through a range of question types that reflect distinct levels of cognitive engagement—from *Information Inquiry* that seek factual knowledge, to *Representation Inquiry* that request alternative forms of prior explanations, to *Metacognitive Inquiry* that evaluate AI-generated content and monitor one's understanding.

A mixed-effects model revealed a significant main effect of question type ($p < .001$), indicating that certain question types were used far more frequently than others. This suggests that students tended to rely on a limited subset of question types, while the majority of types appeared only rarely.

Although the overall interaction between academic competence and question type was not significant, post-hoc analyses (Table 2) revealed marginally significant positive associations for three question types within the Metacognitive Inquiry category; *Critique* ($p = .094$), *Confirmation* ($p = .069$), and *Practice Problem Generation* ($p = .062$). These results suggest that higher-performing students were more likely to engage in metacognitive regulation. Rather than passively accepting GenAI's outputs, they critically evaluated the responses against their growing understanding and identified inconsistencies in the ongoing conversation. They also proactively monitored the accuracy of their understanding by requesting confirmation of unclear points and generating practice problems to assess their mastery.

Table 2: Mixed-effects model results for question types as a function of academic competence. Coefficients are reported relative to Concepts, which serves as the reference category. Statistical significance is denoted as $p < 0.1$ ($^{+}$), $p < 0.05$ (*), $p < 0.01$ (**), or $p < 0.001$ (***).

Question Type	Estimate (b)	SE	p	Sig.
Information Inquiry				
Concepts (Ref.)	-	-	-	-
Methods	-0.160	0.164	.328	.
Reasons	0.311	0.202	.123	.
Representation Inquiry				
Elaboration	-0.100	0.146	.492	.
Rephrasing	0.091	0.188	.629	.
Simplification	-0.184	0.171	.280	.
Material Generation	0.022	0.137	.871	.
Summarization	-0.147	0.213	.490	.
Metacognitive Inquiry				
Self-Explanation	0.271	0.240	.259	.
Confirmation	0.269	0.148	.069	+
Critique	0.391	0.233	.094	+
Practice Problem Gen.	0.347	0.185	.062	+
Extending	0.100	0.298	.736	.
Learning Objectives	0.108	0.137	.430	.
Miscellaneous				
Social Expression	0.313	0.383	.414	.
Translation	0.195	0.184	.287	.

3.2 Better Matecognitive Calibration and Mastery in Higher-Performing Students

3.2.1 Quantitative Comparison of Queries and Their Patterns. High-performing students submitted an average of 13.00 messages ($SD = 5.89$), and low-performing students submitted 11.26 messages ($SD = 4.04$), with no significant difference between the two groups ($p = .25$). In contrast, a t-test revealed a significant difference in average length of questions ($p < .05$): high-performing students wrote longer questions on average ($M = 450.83$, $SD = 309.50$) than low-performing students ($M = 309.09$, $SD = 116.34$). These results suggest that high-performing students, despite asking a similar number of questions, posed longer and more detailed questions.

For the distribution of question types (Figure 1), we observed that high-performing students asked roughly twice as many *Metacognitive Inquiries* as low-performing students. Low-performing students showed a greater proportion of *Representation* and *Information Inquiries*.

3.2.2 Metacognitive Calibration. As shown in Figure 2, a t-test revealed that high-performing students had significantly lower

metacognitive calibration scores ($M = -1.26$, $SD = 1.91$) than low-performing students ($M = 0.39$, $SD = 1.53$; $p < .01$). This indicates that high-performing students exhibited more accurate metacognitive monitoring, suggesting that low-performing students tended to be overconfident, whereas high-performing students were slightly underconfident.

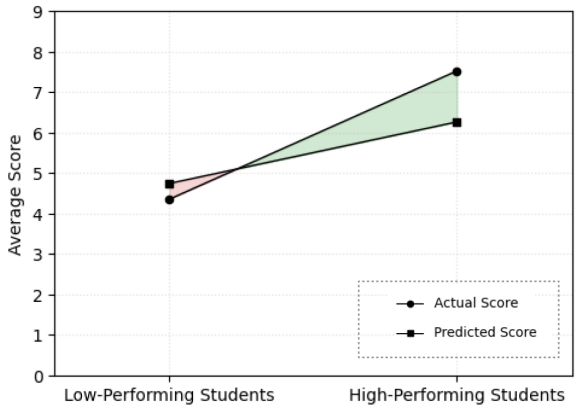

Figure 2: Metacognitive calibration scores for high- and low-performing students. Low-performing students showed a tendency toward overestimation (*predicted* > *actual*), whereas high-performing students exhibited underestimation (*actual* > *predicted*).

3.2.3 Perceived Mastery. A Mann–Whitney U test showed that, after the study, high-performing students reported significantly higher perceived mastery than low-performing students ($p < .05$; Figure 3). This suggests that differences in questioning patterns also shaped students' positive learning experiences.

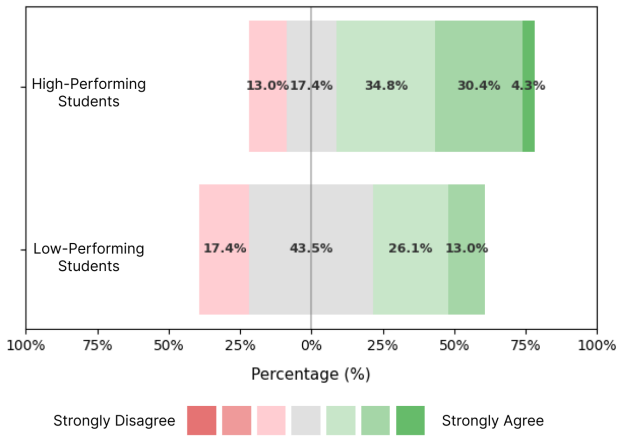

Figure 3: Stacked bar charts of seven-point Likert ratings on perceived mastery for high- and low-performing students.

4 Discussion

Our results suggest that metacognitive regulation is a key underlying factor shaping the still-underexplored notion of "asking good questions to AI." Higher-performing students actively produced more *Metacognitive Inquiries*, revealing frequent monitoring of their own understanding and critical evaluations of the AI's answers. From our results, we observed that students' questioning behaviors serve as a potential indicator of meaningful learning. This implies that question patterns and frequencies—when triangulated with complementary performance indicators—can provide a rich metric for assessing the effectiveness of GenAI in self-directed learning. Future work can extend this investigation by examining the feasibility of metacognitive questioning as a generalizable approach across other diverse educational contexts.

Building on these insights, we identify several implications for designing AI-supported learning systems, especially for learners who struggle to engage in spontaneous metacognitive regulation.

First, literacy efforts could go beyond teaching prompt engineering techniques to include the art of questioning, such as Socratic inquiry, critical reasoning, and metacognitive self-checking. Without this broader skillset, low-performing learners may default to superficial questioning even when interacting with advanced AI systems.

Second, question generation would benefit from remaining primarily within the learner's cognitive space. While fully automated next-question suggestions can improve efficiency, they risk replacing students' own reasoning processes. In educational contexts, AI systems may instead provide templates or partial scaffolds for higher-level questions, enabling students to formulate the final question themselves.

Third, existing frameworks [6, 9] for categorizing questions—often grounded in Bloom's taxonomy [13]—can be expanded by incorporating the metacognitive dimension that emerged in our study. When students repeatedly ask low-level factual questions rather than building toward higher-order inquiry, AI systems may gently nudge them toward deeper forms of questioning that encourage integration, monitoring, and evaluation of knowledge.

Finally, although modern AI systems increasingly rely on knowledge tracing and predictive modeling to identify students' knowledge gaps, proactively supplying information before a learner recognizes the need can undermine metacognitive awareness. Instead of preemptively providing answers, systems can incorporate reflective checkpoints, such as brief quizzes or prompts that ask learners to articulate what they do and do not understand.

Such scaffolding, however, must be applied carefully for high-performing learners, who may be prone to underconfidence [5] and could be hindered by excessive intervention [4]. These findings collectively suggest that AI-infused educational tools may benefit from adapting their scaffolding strategies to the learner's competence level and questioning patterns, rather than adopting a one-size-fits-all approach.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2024-00353125).

References

[1] Fred Biddulph, David Symington, and Roger Osborne. 1986. The place of children's questions in primary science education. *Research in Science & Technological Education* 4, 1 (1986), 77–88.

[2] Eason Chen, Danyang Wang, Luyi Xu, Chen Cao, Xiao Fang, and Jionghao Lin. 2024. A systematic review on prompt engineering in large language models for k-12 stem education. *arXiv preprint arXiv:2410.11123* (2024).

[3] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[4] Xusheng Dai, Zhaochun Wen, Jianxiao Jiang, Huiqin Liu, and Yu Zhang. 2025. How Students Use AI Feedback Matters: Experimental Evidence on Physics Achievement and Autonomy. *arXiv preprint arXiv:2505.08672* (2025).

[5] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*. Vol. 44. Elsevier, 247–296.

[6] Yue Fu and Alexis Hiniker. 2025. Supporting students' reading and cognition with AI. *arXiv preprint arXiv:2504.13900* (2025).

[7] Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal* 31, 1 (1994), 104–137.

[8] Kathleen A Harper, Eugenia Etkina, and Yuhfen Lin. 2003. Encouraging and analyzing student questions in a large physics course: Meaningful patterns for instructors. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 40, 8 (2003), 776–791.

[9] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach ai how to code: Using large language models as teachable agents for programming education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–28.

[10] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the effect of AI code generators on supporting novice learners in introductory programming. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–23.

[11] Shahedul Huq Khandkar. 2009. Open coding. *University of Calgary* 23, 2009 (2009), 2009.

[12] Alison King. 1989. Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology* 14, 4 (1989), 366–381.

[13] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.

[14] Qianou Ma, Kenneth Koedinger, and Tongshuang Wu. 2025. Not Everyone Wins with LLMs: Behavioral Patterns and Pedagogical Implications in AI-assisted Data Analysis. *arXiv preprint arXiv:2509.21890* (2025).

[15] Gili Marbach-Ad and Phillip G Sokolove. 2000. Can undergraduate biology students learn to ask higher level questions? *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 37, 8 (2000), 854–870.

[16] Junho Myung, Hyunseung Lim, Hana Oh, Hyoungwook Jin, Nayeon Kang, So-Yeon Ahn, Hwajung Hong, Alice Oh, and Juho Kim. 2025. When Scaffolding Breaks: Investigating Student Interaction with LLM-Based Writing Support in Real-Time K-12 EFL Classrooms. *arXiv preprint arXiv:2512.05506* (2025).

[17] John L Nietfeld, Li Cao, and Jason W Osborne. 2006. The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and learning* 1, 2 (2006), 159–179.

[18] OpenAI. 2025. ChatGPT-4o. https://openai.com/index/hello-gpt-4o Accessed: Feb 1, 2026.

[19] Nicholas Otis, Rowan Clarke, Solene Delecourt, David Holtz, and Rembrand Koning. 2024. The uneven impact of generative AI on entrepreneurial performance. (2024).

[20] Gregory Schraw. 2009. A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and learning* 4, 1 (2009), 33–45.

[21] Dan Sun, Azzeddine Boudouaia, Junfeng Yang, and Jie Xu. 2024. Investigating students' programming behaviors, interaction qualities and perceptions through prompt-based learning in ChatGPT. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–14.

[22] Mo Wang, Minjuan Wang, Xin Xu, Lanqing Yang, Dunbo Cai, and Minghao Yin. 2023. Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies* 17 (2023), 629–641.

[23] Nance Speizman Wilson and Linda Smetana. 2011. Questioning as thinking: a metacognitive framework to improve comprehension of expository text. *Literacy* 45, 2 (2011), 84–90. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1741-4369.2011.00584.x doi:10.1111/j.1741-4369.2011.00584.x

[24] Yeonsun Yang, Ahyeon Shin, Mincheol Kang, Jiheon Kang, Xu Wang, and Jean Y. Song. 2025. Easy Come, Easy Go? Examining the Perceptions and Learning Effects of LLM-based Chatbot in the Context of Search-as-Learning. arXiv:2410.01396 [cs.HC] https://arxiv.org/abs/2410.01396

[25] J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–21.

[26] Uri Zoller. 1987. The fostering of question-asking capability: A meaningful aspect of problem-solving in chemistry. *Journal of Chemical Education* 64, 6 (1987), 510.

[27] Uri Zoller, Georgios Tsaparlis, Michal Fatsow, and Aviva Lubezky. 1997. Student self-assessment of higher-order cognitive skills in college science teaching. *Journal of College Science Teaching* 27, 2 (1997), 99.