Brains vs. Algorithms? How Experts and Students See AI-Generated Distractors

Zifeng Liu[1], **Hai Li**[1], **Jie Chao**[2], **Wanli Xing**[1]

[1]College of Education, University of Florida, Florida, USA
[2]The Concord Consortium, Concord, Massachusetts, USA
liuzifeng@ufl.edu, li.ha@ufl.edu, jchao@concord.org, wanli.xing@coe.ufl.edu

Abstract

Multiple-choice questions (MCQs) are central to instruction and assessment, with distractors revealing student understanding and misconceptions. However, creating high-quality distractors is time-consuming, especially for emerging domains like K–12 AI education. This study explores using generative AI to support distractor creation in a self-paced online module integrating AI and Algebra 1. Five MCQs were selected to compare distractors written by human developers and ChatGPT, using expert reviews and log data from 80 students. Experts rated human distractors higher overall, though AI ones consistently ranked second. Log analysis showed human distractors drew more initial selections, while students who chose AI distractors spent more time engaging without differences in hint use or revisits. Transition patterns across attempts suggest AI-generated distractors can effectively guide students toward correct answers, highlighting their potential for scalable MCQ design.

Introduction

Multiple-choice questions (MCQs) have long been a widely used format in both instruction and assessment (Bao et al. 2025; Raina, Liusie, and Gales 2023). However, curriculum and assessment developers must devote substantial time and effort to crafting high-quality MCQs (Alhazmi et al. 2024; Raina, Liusie, and Gales 2023), particularly their distractors. A distractor, defined as an incorrect option presented alongside the correct answer (key) and the question stem, plays a critical role in the diagnostic function of MCQs. Despite the considerable time invested in developing distractors, ensuring their quality remains challenging. Human-created distractors are often subjective and may not accurately reflect students' genuine misconceptions (Du, Xing, and Zhang 2022; Fernandez et al. 2024; Shin, Guo, and Gierl 2019). Moreover, in large-scale learning environments, students' misconceptions and knowledge gaps are often difficult to detect and address in a timely manner (Alhazmi et al. 2024; Qian and Lehman 2017). For emerging subjects such as K–12 AI education, where reference materials are limited, these challenges are especially acute (Kumar et al. 2023; Liu et al. 2024a).

The rise of generative AI (GenAI) offers considerable potential for scaling the production of distractors. Prior research has explored the use of models such as GPT for distractor generation (e.g., Doughty et al. 2024; Feng et al. 2024). However, limitations remain. Most existing studies have focused on subjects such as mathematics (e.g., Feng et al. 2024; Fernandez et al. 2024), language learning (e.g., Bitew et al. 2023; Ludewig, Schwerter, and McElvany 2023), reading comprehension (e.g., Gao et al. 2019; Zhou, Luo, and Wu 2020) and higher education level computing or programming courses (e.g., Alhazmi et al. 2024; Doughty et al. 2024), while applications in emerging fields like K–12 AI education remain scarce.

Traditional approaches in educational measurement rely on psychometric indicators and expert review to assess distractor quality (DiBattista and Kurzawa 2011; Haladyna and Rodriguez 2013). While these methods provide standardized criteria, they face constraints in dynamic online learning environments and emerging subject domains (Liusie et al. 2023; Mulla and Gharpure 2023). With the development of educational technology and learning analytics, researchers have increasingly turned to behavioral log data to gain new insights into students' cognitive processes and strategies (Arizmendi et al. 2023; Liu et al. 2024b; Monteith et al. 2024); however, few studies have used log data to analyze how students interact with different distractors. Moreover, there remains a lack of specialized evaluation approaches for AI-generated distractors (Alhazmi et al. 2024; Liusie et al. 2023), and empirical evidence in the context of K–12 AI education is particularly limited.

The motivation of this study lies in responding to these challenges. Building on a large-scale online K–12 AI education curriculum, we recruited four expert reviewers and more than 80 students. During the online learning module, students encountered five MCQs containing both AI-generated and human-generated distractors. We collected experts' selections of the most plausible distractors along with their qualitative comments, as well as detailed log data capturing students' interactions during problem-solving. To guide our study, we pose the following three research questions (RQs):

- **RQ1**: To what extent do experts prefer AI- over human-generated distractors in MCQs?
- **RQ2**: To what extent are students more likely to select AI- rather than human-generated distractors in MCQs?

- **RQ3**: How is distractor type (AI- vs. human-generated) associated with students' behavioral engagement, including time spent, hint use, and content revisiting?

Related Work

Human vs. AI-Generated MCQ Distractors

Human-generated distractors have traditionally been the predominant approach in assessment design, valued for their potential to capture student misunderstandings through educators' subject knowledge and teaching experience (Haladyna and Rodriguez 2013). However, the process of creating human distractors is inherently labor-intensive, making it difficult to ensure both efficiency and consistency of quality in large-scale curriculum development (Doughty et al. 2024; Raina, Liusie, and Gales 2023), particularly when expert resources are scarce in emerging fields such as K–12 AI education.

Recent studies explored GenAI to generted distractors and indicate that AI-generated distractors perform well in terms of linguistic fluency and surface plausibility and can be comparable to human-generated ones in certain standardized tests (Awalurahman, Aji, and Budi 2025; Kurdi et al. 2020). However, in-depth analyses also reveal key limitations: AI-generated distractors often lack a deep understanding of domain-specific concepts and struggle to accurately reflect students' typical errors in complex reasoning (Alhazmi et al. 2024). To increase the efficiency and quality of the GenAI-generated distractors, more studies still needed for better understanding of how AI-generated distractor work or complemente or short to human-created ones.

Existing comparative studies predominantly rely on expert subjective evaluation or small-scale experiments (Bitew et al. 2023; Doughty et al. 2024; Wang, Xiao, and Tseng 2025), with few incorporating empirical evidence based on large-scale authentic student interaction data. To address this gap, the present study provides a comprehensive comparison of human- and AI-generated distractors in K–12 AI education by combining expert assessment with student log data, thereby revealing their performance differences in real learning environments.

Evaluation of MCQ Distractors

The evaluation of MCQ distractors traditionally relying on psychometric indicators (e.g., attractiveness, discrimination index, difficulty coefficient) and expert review to quantify their effectiveness and content validity (DiBattista and Kurzawa 2011; Haladyna and Rodriguez 2013). While these methods provide standardized evaluation criteria, their applicability and timeliness are limited in dynamic online learning environments and emerging subject domains. Distractor evaluation faces well-recognized challenges, with prior studies typically restricted to small sets of MCQs or lacking systematic evaluation (Rodriguez-Torrealba, Garcia-Lopez, and Garcia-Cabot 2022; Vachev et al. 2022).

With the development of educational technology and learning analytics, evaluation methods have become increasingly diversified. Learning analytics techniques leverage student behavioral logs to provide new perspectives for evaluation, revealing students' cognitive processes and learning strategies when interacting with different distractors, such as response time, hint usage, and content review behaviors (Bergdahl et al. 2024; Liu, Ngo, and Xing 2025).

Despite these advances, specialized approaches for effectively evaluating AI-generated distractors remain underdeveloped, and empirical applications in emerging fields such as K–12 AI education are scarce. This study addresses this gap by constructing a multi-dimensional evaluation that integrates both qualitative and quantitative analyses through expert assessment and student log data, systematically evaluating the pedagogical impact of GenAI-generated distractors in K–12 AI education.

Method

Study Context

To address the RQs, we selected five MCQs (see Table 1) from an online AI education curriculum as the basis for exploratory analysis. The curriculum is delivered in a virtual learning environment that allows high school students to learn at their own pace while engaging with new and emerging topics in AI. Specifically, the curriculum is designed to help students understand the relationship between text analysis and algebraic equations, culminating in the development of a functional sentiment analysis model.

The curriculum consists of five sequential learning activities, progressing from introductory concepts to more advanced applications, and spans approximately 250 minutes in total. MCQs are embedded throughout to assess comprehension and reinforce key concepts, serving not only to evaluate content knowledge but also to provide immediate, substantive feedback.

Data Description

MCQs The five selected MCQs were drawn from Activity 1 and Activity 2 of the curriculum. Each MCQ includes a question stem, one or more correct answers (keys), and a set of distractors generated either by human developers or by ChatGPT. Table 1 summarizes the stems, correct answers, and both human- and AI-generated distractors for the MCQs. Figure **??** illustrates the overall structure of these questions. Each MCQ consists of two components: a contextual information (presented through images or text) and a corresponding MCQ item. Students engage with the contextual material before answering the question, and then receive automated feedback (i.e., hints) after submitting their response by clicking the "Check Answer" button.

Distractor Generation We used ChatGPT 4o to generate distractors for MCQs, building on recent work with GPT models (Doughty et al. 2024). A zero-shot approach was employed with a standardized prompt containing the context, question stem, correct answer (key), and explicit instructional constraints.

Distractors were validated by removing empty or faulty outputs, with regeneration as needed. A human reviewer further refined the results by eliminating irrelevant content. The final set of distractors, including overlaps with teacher-created options, was compiled for analysis.

#	Question stem	Correct answer (feedback)	Distractors (feedback)
1	Which of the following are outputs from this model? Select all that apply.	Positive; Negative (Yes! You are correct.)	1*. Emotion Type (Good thinking. But refer to the diagram above—what are the two defined outputs?) 2. Neutral (Good thinking. Indeed, there are relatively neutral reviews, but this model has been defined to only generate two outputs: positive or negative.) 3. Mixed (It is true that some reviews are mixed, but this model has been defined to only generate two outputs: positive or negative.) 4*. Text Review (Incorrect. "Text Review" is the input. Look closely at the model again.)
2	Which of the following can Mini find in the review to the left? Check all that apply.	wow; !!!; worth the price; :-); ?! (Yes! You are correct.)	1. Top notch (Look again. There is a hyphen in "top-notch".) 2*. Not bad (Look again—there is no "Not bad" in the review.) 3*. Too expensive (Sorry, the review does not mention "too expensive".) 4*. Amazing deal (Be careful. Mini can only find exact words, not overall impressions.) 5. Wait (Be careful. The word in the review is "waited", not "wait".)
3	How is the true sentiment of a review determined?	The true sentiment is how a person would categorize or label a review (Yes! You are correct.)	1. The true sentiment is how Mini would categorize or label a review (Sorry, that is incorrect.) 2*. The true sentiment is the most common word used in the review (Not quite. Word choice plays a role, but sentiment isn't based on frequency.) 3. A sentiment analysis model (Sorry, that is incorrect.) 4*. The true sentiment is automatically calculated based on the length of the review (Incorrect.)
4	What does the Cbest variable represent in the Algebraic Expression?	The count of the word "best" (Correct! The count of "best" is represented by the variable "Cbest".)	1*. The number of positive words in the review (Not quite. "Cbest" only counts "best".) 2*. The total number of words in the review (That's not it.) 3. The input of the Sentiment Analysis Model (The input is the text of the review.) 4. The output of the Sentiment Analysis Model (The output is positive or negative.) 5*. The overall sentiment score of the review (Good try! But "Cbest" is not the score itself.) 6. The true sentiment of the review (The true sentiment can only be determined by a human.)
5	What about Review F: "Their pizza is definitely some of the best around, and the white pizza is one of the best items on the menu." (True sentiment: positive.) What are the results of the model?	Mini doesn't have a rule for this case. (Yes, you are correct!)	1. Cbest = 0 therefore S = positive (Incorrect! Count again.) 2*. Cbest = 0 therefore S = negative (Incorrect! Count again.) 3. Cbest = 1 therefore S = negative (Incorrect! Count again.) 4*. Cbest = 2 therefore S = positive (Incorrect! Does Mini have a rule?) 5*. Cbest = 2 therefore S = negative (Incorrect! Does Mini have a rule?)

Table 1: Five questions used in this study. *Note*: * means distractor from GenAI. For Q5, AI and human both have distractor 4.

Expert Review To address RQ1, four experts were invited to review all the distractors associated with five selected questions. The panel consisted of a learning scientist, and three educational researchers, all with extensive experience in K–12 AI education.

Each expert was provided with the full context of each question, including the contextual information, the correct answer(s), all distractors, and the feedback associated with each distractor. To ensure unbiased evaluation, the source of each distractor (i.e., AI-generated or human-written) was not disclosed. For each question, experts were asked to select the two or three distractors they considered most appropriate, based on qualities such as relevance, plausibility, and alignment with the question's learning goals. These selections reflected their overall judgment of the question's quality. In addition, experts were required to provide written comments on each distractor, explaining why it was considered appropriate or inappropriate with respect to the question stem and intended learning outcomes.

Student Interaction Logs To address RQ2 and RQ3, we analyzed log data collected from 80 high school students who self-selected to participate in the online AI education curriculum. These students completed the full set of activities without being informed which items contained AI-generated distractors. The system automatically recorded their interactions with each MCQ, capturing data on their selected options and the timing of key actions.

In total, the five MCQs yielded a total number of 13,772 log entries. The system captures student behavior through a series of structured events. For example, the *event* field specifies the type of action taken (e.g., *answer_selected*, *check_answer*, *focus_in*). The logs include detailed records of when students selected each distractor (*answer_selected*) and when they clicked the "Check Answer" button (*check_answer*). These time-stamped events enable analysis of distractor selection behavior, timing patterns, and engagement indicators such as time spent on each question, hint usage, and content revisiting.

Data Analysis

Qualitative Analysis The qualitative analysis focused on expert reviews of each distractor to gain insights into their judgments about what constitutes a high-quality distractor in the context of K-12 AI education. Experts provided written comments for every distractor, regardless of whether it was selected as one of their top choices. These comments were analyzed using inductive thematic coding, which involved identifying recurring themes and evaluative criteria mentioned by the experts. The coding process aimed to capture both positive and negative rationales, such as semantic relevance, plausibility, alignment with the question stem, and cognitive challenge level.

Quantitative Analysis The quantitative analysis consisted of two components: (1) analyzing expert selections of the most appropriate distractors for each MCQ, and (2) analyzing student interaction logs to examine behavioral engagement patterns associated with AI-generated versus human-written distractors. For expert selections, we calculated how often each distractor was chosen among the top two or three options. For student behavior analysis, we examined three key engagement indicators: (1) time spent on each question, measured as the duration between the student's first recorded interaction (e.g., *answer_selected*) and their final event on that question; (2) hint usage, operationalized as the occurrence of a *check_answer* event, indicating that the student actively sought feedback; and (3) content revisiting, determined by whether the student re-engaged with related contextual material while working on the question.

Results

Expert Evaluation (RQ1)

Figure 1 (a) summarizes expert selections of the top distractors across the five MCQs, with distractors color-coded as either AI-generated (blue) or human-written (green). For Questions 1 to 3, each expert was instructed to select two top distractors. For Questions 4 and 5, each expert selected three top distractors from the full set of available options.

Among the 12 total top distractors ultimately selected across all questions (i.e., distractors receiving the highest number of expert votes), human-generated distractors were selected 7 times, while AI-generated distractors were selected 5 times, indicating a slightly stronger overall preference for human-written distractors. However, certain AI-generated distractors such as D5 in both Q4 and Q5, received strong agreement from multiple experts, demonstrating that GenAI can produce distractors of competitive quality in this specific educational contexts.

When analyzing the most frequently selected distractor per question, human-generated options received the highest number of votes in Q1, Q3, and Q5, with AI-generated distractors consistently ranking second. In contrast, for Q4, an AI-generated distractor received the highest number of expert votes, surpassing any human-written alternative.

While human-generated distractors were more frequently selected as top choices, the overall vote distribution was relatively balanced. Across all expert selections, AI-generated distractors received a total of 22 votes, while human-generated distractors received 28, yielding a ratio of approximately 11:14. This near parity suggests that AI-generated distractors were generally considered competitive and frequently perceived as high-quality by experts, even if slightly less dominant in final rankings.

In their qualitative evaluation of AI-generated distractors, experts provided detailed comments highlighting both the strengths and limitations of these items across different questions. In Question 2, where AI-generated distractors were generally not preferred, experts emphasized that effective distractors should exhibit explicit or clearly implied connections to the source text. For example, distractors such as "Not bad" and "Too expensive" were criticized for lacking both literal and semantic alignment with the review, thereby limiting their diagnostic value. In contrast, "Amazing deal" was recognized for its emotional congruence with the review's tone, even though the phrase was not directly quoted. This contrast suggests that experts value a combination of textual relevance and cognitive plausibility when

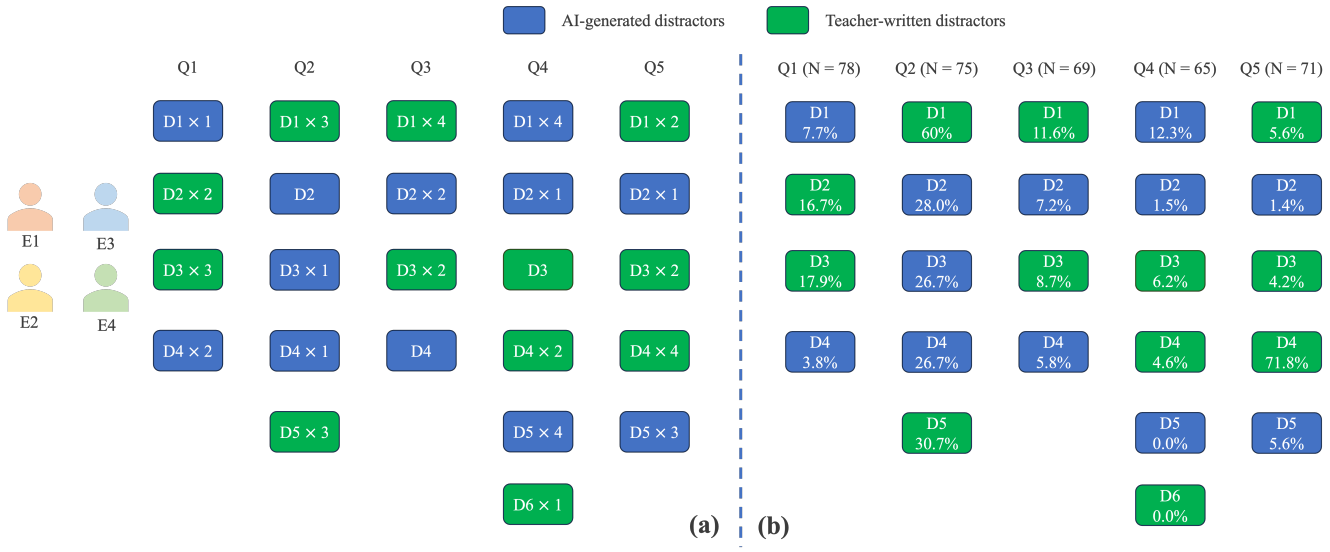

Figure 1: (a) Expert selections of AI- vs. human-generated distractors across five MCQs; (b) Student first-attempt choices of AI- vs. human-generated distractors across five MCQs. *Note*: D1 refers to Distractor 1 in Table 1. "$D1 \times 4$" indicates it was selected four times as a top distractor by the experts. $N = 78$ denotes the number of students who completed the question and for whom log data were collected.

evaluating distractor quality.

In Question 4, expert feedback identified three common reasoning errors embedded in the AI-generated distractors, which were considered useful for assessing students' conceptual understanding. First, D1 overgeneralized the meaning of the variable Cbest, interpreting it as the count of all positive words rather than the specific occurrence of "best." Second, D2 incorrectly equated Cbest with the total number of words in the review, a connection experts considered plausible and sufficiently misleading. Third, D5 suggested that Cbest represented an overall sentiment score or emotional rating. Although inaccurate, experts noted that this interpretation reflected a common student misconception and thus functioned as an effective distractor for assessing students' understanding of the model's variable definitions.

Student Selection Patterns (RQ2)

Figure 1 (b) illustrates students' first-attempt selections for each distractor option across five MCQs. Each cell indicates the percentage of students who selected a given distractor, based on the total number of responses for that question.

Across all five questions, human-written distractors tended to attract higher selection rates. For instance, in Q2, distractor D1 (by a human developer) was chosen by 60% of students. Similarly, in Q5, human-written distractor D4 drew the highest selection rate at 71.8%. These results suggest that human-written distractors were generally more effective in drawing student attention, likely due to their alignment with instructional context.

AI-generated distractors also demonstrated the ability to confuse students in certain contexts. Notably, in Q2 (a "check all that apply" question), AI-generated distractors D2, D3, and D4 were each selected by approximately 26.7–28% of students, indicating their competitive plausibility. However, in other questions, AI-generated distractors were less frequently chosen. The only AI-generated distractor with a selection rate above 10% outside Q2 was D1 in Q4 (12.3%).

Interestingly, students' selection patterns largely aligned with expert preferences, with human-written distractors generally ranking higher among both groups. A notable exception occurred in Q4: distractor D5 received the most expert votes but was not selected by any student at the first attempt. This discrepancy suggests a potential misalignment between expert judgment and student interpretation for that item.

First-Attempt Distractor Selection and Learning Behaviors (RQ3)

Table 2 presents the descriptive statistics and results of the Mann–Whitney U tests for the five MCQs, comparing students who selected AI-generated distractors (AI group) on their first attempt with those who selected human-generated distractors (Human group). The Mann–Whitney U test was used to determine whether the distributions of the two groups differed significantly on each behavioral indicator (i.e., time on task, hint usage, and revisiting behavior).

For time on task, Students in the AI group spent significantly more time on the items than those in the Human group ($U = 1400, p = 0.011$). The corresponding effect size, Cliff's $\delta = 0.33$, indicates a small-to-medium magnitude of difference. For hint usage, the analysis revealed no significant differences between the AI and Human groups ($U = 1063, p = 0.95$). Cliff's $\delta \approx 0$ further suggests a negligible effect. Revisiting behavior result was similar to hint usage, no significant differences were found between the two groups ($U = 888, p = 0.23$). The effect size was

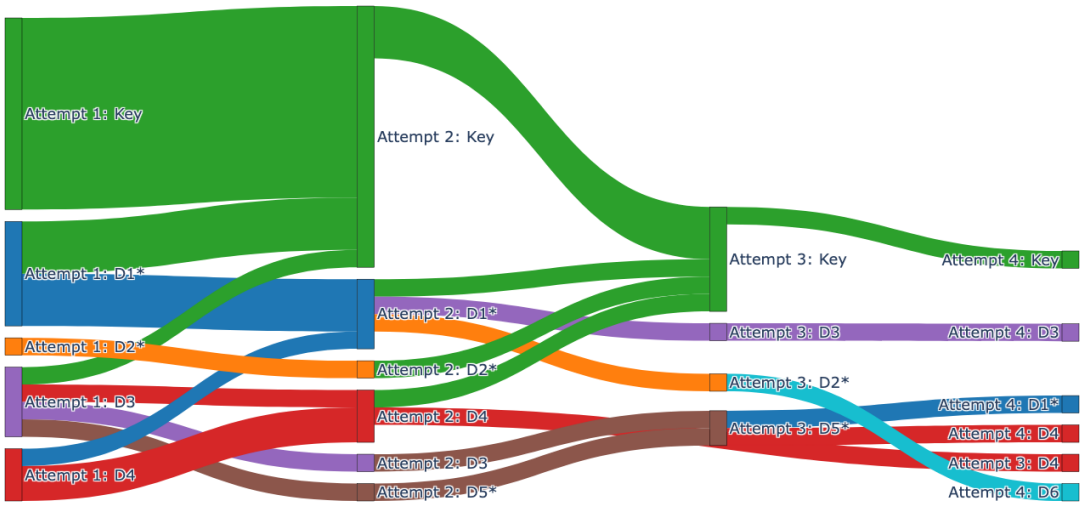

Figure 2: Sankey diagram of student distractor transitions across the first four attempts (Q4). *Note*: * indicates an AI-generated distractor; D1 refers to Distractor 1 in Table 1; Key denotes the correct answer.

Measure	Group	Mean	SD	U	p-value	Cliff
Time (s)	AI Group	420.5	120.3	1400	0.011	0.330
	Human Group	198.7	95.4			
Hints	AI Group	3.1	1.2	1063	0.946	0.009
	Human Group	3.0	1.0			
Revisits	AI Group	10.2	3.4	888	0.227	-0.157
	Human Group	10.8	3.8			

Table 2: Comparison of student learning behaviors.

small (Cliff's $\delta = 0.11$), indicating that the groups exhibited comparable patterns in this behavior.

We further examined how students' choices evolved across different attempts. Figure 2 visualizes student transitions between different option selections across the first four attempts on Q4, using a Sankey diagram to illustrate shifts in conceptual understanding. Each node represents a specific option (one correct answer and four distractors) chosen during a given attempt, while the links indicate how many students moved from one option to another across successive attempts. Among those who initially selected "the number of positive words in the review" (an AI-generated distractor), half arrived at the correct answer on their second attempt. For students who first chose "the total number of words in the review" (another AI-generated distractor), all reached the correct answer by the third attempt. In contrast, students who selected human-created distractors showed more dispersed patterns in subsequent attempts. For instance, students who initially selected "the input of the sentiment analysis model" displayed varied responses in the second attempt, distributed across multiple distractor options.

Discussion

Our findings indicate that experts showed a slight preference for human generated distractors, though AI-generated distractors were often rated comparably and occasionally outperformed human options. This aligns with prior work suggesting that while human developers can more effectively capture students' typical misconceptions due to their pedagogical expertise (Haladyna and Rodriguez 2013), AI-generated distractors can still achieve competitive quality, particularly in surface plausibility and linguistic fluency (Doughty et al. 2024; Feng et al. 2024). The expert comments also revealed detailed criteria in evaluating distractors: textual relevance, semantic plausibility, and alignment with common misconceptions were consistently emphasized. For example, AI-generated distractors in Q4 successfully mirrored typical reasoning errors, which experts judged as valuable for diagnostic assessment. These findings suggest that GenAI can complement human developer expertise by efficiently producing distractors that capture recurring error patterns, though refinement is still needed to ensure contextual fidelity.

Student responses largely paralleled expert judgments, with human-generated distractors more frequently chosen. This is consistent with research showing that distractors grounded in authentic student misconceptions are more attractive to learners (DiBattista and Kurzawa 2011). Nevertheless, AI-generated distractors also attracted a meaningful proportion of students, particularly in multi-answer questions (e.g., Q2), indicating that they can be sufficiently deceptive in specific contexts. The discrepancy observed in Q4, where an AI-generated distractor was highly rated by experts but seldom chosen by students highlights the impor-

tance of triangulating expert review with authentic learner data. Experts may recognize cognitive plausibility in distractors that, in practice, students overlook or interpret differently (Ludewig, Schwerter, and McElvany 2023). This underscores the value of combining expert-based and learner-based evaluation frameworks to assess distractor quality.

Our behavioral analysis showed that students who selected AI-generated distractors spent significantly more time on the items compared to those who chose human-generated distractors. One possible explanation is that AI-generated distractors, while plausible, may be less transparently tied to students' pre-existing misconceptions, requiring additional cognitive effort for students to reconcile their reasoning with the feedback provided. This finding aligns with cognitive load theory, suggesting that less familiar or less contextually grounded stimuli can increase processing demands (Kim et al. 2024). In contrast, no significant group differences emerged in hint usage or revisiting behaviors. This suggests that distractor type may not directly affect students' reliance on external support or their persistence strategies, but it does appear to influence their time investment in problem-solving. The Sankey analysis of Q4 further revealed that students who initially selected AI-generated distractors potentially tended to converge on the correct answer within fewer attempts, whereas those starting with human-generated distractors showed more dispersed trajectories. This may indicate that AI distractors prompt students to engage in more linear reasoning correction, while human distractors induce broader exploration of alternatives.

In summary, these findings provide both theoretical and practical contributions. From a practical perspective, this study makes three key contributions: (1) Improving efficiency: Leveraging AI to automatically generate and analyze distractors at scale can reduce the time and cost of manual design. (2) Identifying gaps: By analyzing student choice patterns, our approach helps developers identify weaknesses in learning materials and uncover typical student misconceptions. (3) Supporting new material development: For emerging domains such as AI education, this method can assist developers in designing more targeted instructional resources and assessments for K–12 learners. From an academic perspective, this study contributes to ongoing discussions in educational technology and computer science education on "automatic AI-assisted instruction." It also extends the literature on automated assessment tools, learning analytics, and the application of GenAI in education.

Conclusion, Limitations and Future Work

This study examined the use of GenAI for distractor generation in the context of a virtual, self-paced AI curriculum for high school students. By combining expert evaluations with analyses of student log data, we provided a multi-faceted view of distractor quality and its potential impact on learning. The findings highlight both the promise and the limitations of AI-generated distractors: although there remains room for improvement, as they did not outperform human-created ones, AI-generated distractors can produce plausible alternatives that align with expert standards, support scalable distractor production, and encourage students to spend more time engaging with the MCQs.

Several limitations should be noted. The study was conducted within a single curriculum and involved a limited set of items, constraining generalizability. It also did not account for variability across AI models or human authors, and behavioral traces alone may not fully reflect students' cognitive processes. Future research could examine distractors across broader subjects and AI models, integrate qualitative methods (e.g., interviews or think-alouds) to capture cognitive reasoning, and develop adaptive systems that use AI-generated distractors for real-time formative feedback to address student misconceptions.

Acknowledgments

This work was supported by a grant from the U.S. Department of Education (Grant No. S411C230070). Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the funding agency.

References

Alhazmi, E.; Sheng, Q. Z.; Zhang, W. E.; Zaib, M.; and Alhazmi, A. 2024. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *arXiv preprint arXiv:2402.01512*.

Arizmendi, C. J.; Bernacki, M. L.; Raković, M.; Plumley, R. D.; Urban, C. J.; Panter, A. T.; and Gates, K. M. 2023. Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, 55(6): 3026–3054.

Awalurahman, H. W.; Aji, R. F.; and Budi, I. 2025. Transformer and Large Language Models for Automatic Multiple-Choice Question Generation: A Systematic Literature Review. *IEEE Access*.

Bao, Q.; Leinonen, J.; Peng, A. Y.; Zhong, W.; Gendron, G.; Pistotti, T.; Huang, A.; Denny, P.; Witbrock, M.; and Liu, J. 2025. Exploring iterative enhancement for improving learnersourced multiple-choice question explanations with large language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2025)*.

Bergdahl, N.; Bond, M.; Sjöberg, J.; Dougherty, M.; and Oxley, E. 2024. Unpacking student engagement in higher education learning analytics: a systematic review. *International Journal of Educational Technology in Higher Education*, 21(1): 63.

Bitew, S. K.; Deleu, J.; Develder, C.; and Demeester, T. 2023. Distractor generation for multiple-choice questions with predictive prompting and large language models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 48–63. Springer.

DiBattista, D.; and Kurzawa, L. 2011. Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, 2(2): 4.

Doughty, J.; Wan, Z.; Bompelli, A.; Qayum, J.; Wang, T.; Zhang, J.; Zheng, Y.; Doyle, A.; Sridhar, P.; Agarwal, A.;

et al. 2024. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference*, 114–123.

Du, H.; Xing, W.; and Zhang, Y. 2022. Misconception of Abstraction: When to Use an Example and When to Use a Variable? In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2*, ICER '22, 28–29. ISBN 9781450391955.

Feng, W.; Lee, J.; McNichols, H.; Scarlatos, A.; Smith, D.; Woodhead, S.; Ornelas, N. O.; and Lan, A. 2024. Exploring automated distractor generation for math multiple-choice questions via large language models. *arXiv preprint arXiv:2404.02124*.

Fernandez, N.; Scarlatos, A.; Feng, W.; Woodhead, S.; and Lan, A. 2024. DiVERT: Distractor generation with variational errors represented as text for math multiple-choice questions. *arXiv preprint arXiv:2406.19356*.

Gao, Y.; Bing, L.; Li, P.; King, I.; and Lyu, M. R. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6423–6430.

Haladyna, T. M.; and Rodriguez, M. C. 2013. *Developing and validating test items*. Routledge.

Kim, M.; Duncan, C.; Yip, S.; and Sankey, D. 2024. Beyond the theoretical and pedagogical constraints of cognitive load theory, and towards a new cognitive philosophy in education. *Educational Philosophy and Theory*, 57(7): 662–673.

Kumar, A. P.; Nayak, A.; Shenoy, M.; Goyal, S.; et al. 2023. A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, 225: 120022.

Kurdi, G.; Leo, J.; Parsia, B.; Sattler, U.; and Al-Emari, S. 2020. A systematic review of automatic question generation for educational purposes. *International journal of artificial intelligence in education*, 30(1): 121–204.

Liu, Z.; Guo, R.; Jiao, X.; Gao, X.; Oh, H.; and Xing, W. 2024a. How AI Assisted K-12 Computer Science Education: A Systematic Review. In *Proceedings of the 2024 ASEE Annual Conference & Exposition*. Portland, OR, United States. Paper presented at the 2024 ASEE Annual Conference & Exposition.

Liu, Z.; Jiao, X.; Li, C.; and Xing, W. 2024b. Fair Prediction of Students' Summative Performance Changes Using Online Learning Behavior Data. In PaaÃŸen, B.; and Epp, C. D., eds., *Proceedings of the 17th International Conference on Educational Data Mining*, 686–691. Atlanta, Georgia, USA: International Educational Data Mining Society. ISBN 978-1-7336736-5-5.

Liu, Z.; Ngo, B.; and Xing, W. 2025. Evaluating AI-Generated Distractors in Programming Education: A Human-AI Collaborative Approach. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V.2*, ICER '25, 12. ISBN 9798400713415.

Liusie, A.; Raina, V.; Mullooly, A.; Knill, K.; and Gales, M. J. F. 2023. Analysis of the Cambridge multiple-choice questions reading dataset with a focus on candidate response distribution. arXiv preprint. arXiv:2306.13047.

Ludewig, U.; Schwerter, J.; and McElvany, N. 2023. The Features of Plausible but Incorrect Options: Distractor Plausibility in Synonym-Based Vocabulary Tests. *Journal of Psychoeducational Assessment*, 41(7): 711–731.

Monteith, B.; Liu, Z.; Chao, J.; et al. 2024. Using Entropy Analysis to Explore Student Engagement in an Online High School Data Science Course. TechRxiv. Preprint.

Mulla, N.; and Gharpure, P. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12: 1–32.

Qian, Y.; and Lehman, J. 2017. Students' misconceptions and other difficulties in introductory programming: A literature review. *ACM Transactions on Computing Education (TOCE)*, 18(1): 1–24.

Raina, V.; Liusie, A.; and Gales, M. 2023. Assessing Distractors in Multiple-Choice Tests. *arXiv preprint arXiv:2311.04554*.

Rodriguez-Torrealba, R.; Garcia-Lopez, E.; and Garcia-Cabot, A. 2022. End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. *Expert Systems with Applications*, 208: 118258.

Shin, J.; Guo, Q.; and Gierl, M. J. 2019. Multiple-choice item distractor development using topic modeling approaches. *Frontiers in psychology*, 10: 825.

Vachev, K.; Hardalov, M.; Karadzhov, G.; Georgiev, G.; Koychev, I.; and Nakov, P. 2022. Leaf: Multiple-choice question generation. In *European Conference on Information Retrieval*, 321–328. Springer.

Wang, J.; Xiao, R.; and Tseng, Y. J. 2025. Generating AI Literacy MCQs: A Multi-Agent LLM Approach. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, 1651–1652. ACM.

Zhou, X.; Luo, S.; and Wu, Y. 2020. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9725–9732.