Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Human-AI Collaborative Assessment of AI-Generated vs. Human-Created MCQ Distractors

Zifeng LIU[a*], Priyadharshini Ganapathy PRASAD[a], Bach NGO[b], Xinyue JIAO[c] & Wanli XING[a]

[a]*University of Florida, the United States*
[b]*The Frazer School, the United States*
[c] *New York University, the United States*
*liuzifeng@ufl.edu

Abstract: Multiple-choice questions (MCQs) are a widely used and effective assessment method, with the quality of distractors being crucial for their effectiveness. Recent studies have explored the use of large language models (LLMs) to generate distractors. However, generating high-quality distractors remains challenging in subjects such as computer science and mathematics that requiring strong reasoning. While some research has investigated AI-generated distractors in these fields, evaluating their quality is difficult due to existing metrics primarily focusing on surface semantics and failing to capture the necessary reasoning. This study introduces a human-AI collaborative assessment approach to evaluate distractor quality. We applied this method to compare AI-generated and human-created distractors in two high school courses: programming (N = 349) and statistics (N = 576). The findings suggest that AI-generated distractors can be competitive with human-created ones in programming courses, but significant differences exist in the understand, analyze, and evaluate types of MCQs in statistics ($p < 0.01$). This study provides a practical and scalable solution for integrating and evaluating AI-generated distractors in educational assessments.

Keywords: Distractor, generative AI, evaluation, human-AI collaboration, multi-choice question

1. Introduction

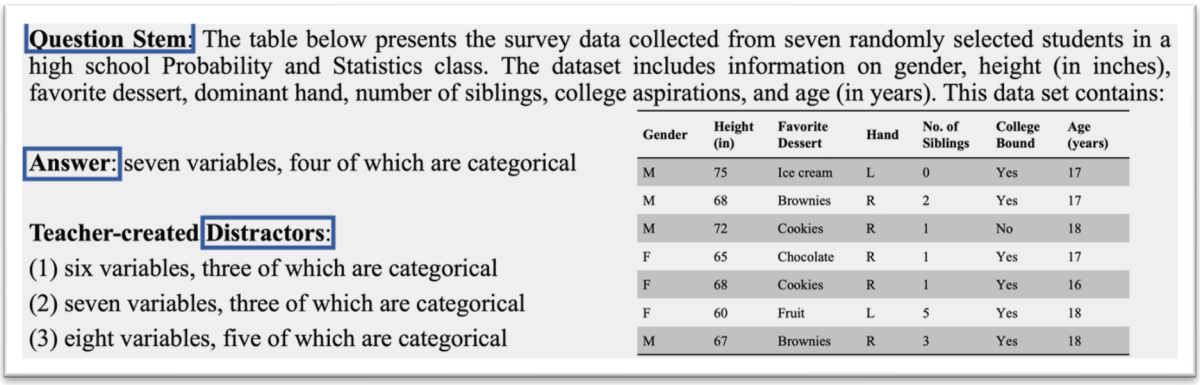
The rapid advancement of generative artificial intelligence (AI) has revolutionized assessments in education, where AI-driven approaches have been increasingly applied to automate generate test questions (Rodrigues et al., 2024) or distractors in multi-choice questions (MCQ) (Doughty et al., 2024). Recent AI methods for distractor generation include deep neural networks (e.g., Zhou et al., (2020)) and pre-trained language models (e.g., Bulathwela et al., (2023)). However, these methods have predominantly focused on generating distractors for reading comprehension and language learning assessments using techniques such as fine-tuning and prompting.

Recent research has explored various neural approaches to generating contextually appropriate multiple-choice question (MCQ) distractors. For example, the BERT-based Distractor Generation (BDG) model leverages multi-task learning and negative answer training to produce diverse and instructionally meaningful distractors (Chung et al., 2020). Building on this line of work, T5-based models have been adapted using techniques such as closed-book question answering and binary classification to improve the semantic relevance of distractors (Lelkes et al., 2021). Beyond fine-tuning, prompting has emerged as a lightweight yet powerful alternative, enabling pre-trained language models to generate distractors through textual instructions without the need for additional labeled data. Large language models (LLMs) like GPT-3, GPT-4, and ChatGPT have shown strong performance in this regard, using strategies such as zero-shot, few-shot, single-stage, and multi-stage prompting (Bitew et al., 2023;

McNicholas et al., 2023). For instance, zero-shot prompting involves providing the model with only the question stem and correct answer, asking it to generate plausible incorrect options. Few-shot prompting enhances the quality of distractors by incorporating several annotated examples within the prompt. In addition, chain-of-thought (CoT) prompting has been used to guide LLMs in simulating common student misconceptions, thereby improving the pedagogical effectiveness of generated distractors (Wei et al., 2022).

Despite the success of fine-tuned and prompted models in generating high-quality distractors, challenges remain. While AI-generated distractors can significantly reduce educators' workload by serving as an initial draft, human review and refinement remain necessary to ensure their pedagogical effectiveness. Evaluating the quality of AI-generated distractors is an identified challenge of this research area. Distractor evaluation methods are typically categorized into automatic and human evaluation techniques. Automatic evaluation employs metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to assess similarity between generated and human-crafted distractors. Ranking-based and n-gram-based metrics further quantify the quality and relevance of distractors by analyzing their position within a ranked list or their lexical overlap with ground truth distractors (Lavie & Denkowski, 2009; Papineni et al., 2002). While these metrics offer scalable evaluation, they may not fully capture the pedagogical effectiveness of distractors.

To address these challenges, this study proposes a human-AI collaborative assessment framework to compare distractor quality from AI and from human teachers. We apply this framework to programming and statistics courses by comparing AI-generated distractors with human-created ones, enabling a more structured and scalable evaluation process. Specifically, our approach integrates fine-tuning and prompting-based AI models, explores advanced evaluation metrics that extend beyond surface-level similarity, and incorporates human-in-the-loop validation to improve both the accuracy and pedagogical value of AI-generated distractors. By systematically assessing the effectiveness of different AI-assisted techniques in MCQ design, this study contributes to a deeper understanding of how AI can enhance educational assessments.



Question Stem: The table below presents the survey data collected from seven randomly selected students in a high school Probability and Statistics class. The dataset includes information on gender, height (in inches), favorite dessert, dominant hand, number of siblings, college aspirations, and age (in years). This data set contains:

Gender	Height (in)	Favorite Dessert	Hand	No. of Siblings	College Bound	Age (years)
M	75	Ice cream	L	0	Yes	17
M	68	Brownies	R	2	Yes	17
M	72	Cookies	R	1	No	18
F	65	Chocolate	R	1	Yes	17
F	68	Cookies	R	1	Yes	16
F	60	Fruit	L	5	Yes	18
M	67	Brownies	R	3	Yes	18

Answer: seven variables, four of which are categorical

Teacher-created Distractors:

(1) six variables, three of which are categorical

(2) seven variables, three of which are categorical

(3) eight variables, five of which are categorical

Figure 1 An Example of the collected data

2. Method

2.1 Data Description

Data was collected on Florida Virtual School from two high school courses: Foundations of Programming and Foundations of Statistics. The programming course contains 349 MCQs, while the statistics course includes 576 questions. Typically, in this dataset, each MCQ consists of a question stem with textual supporting materials (e.g., tables), a correct answer, and three distractors manually created by teachers, shown in Figure 1.

Table 1. The Bloom's taxonomy levels of the MCQs in the dataset

Course	R	U	App	Ana	E	C	Total

Programming	91	110	87	40	18	3	349
Statistics	4	43	185	269	71	4	576
Total	95	153	272	309	89	7	925

R: Remember, U: Understand, App: Application, Ana: Analysis, E: Evaluation, C: Create

2.2 Bloom's Taxonomy Levels

To investigate AI's ability to generate distractors for different types of questions, we categorized all the MCQs (N = 925) based on Bloom's taxonomy (Krathwohl, 2002). All MCQs were classified into six cognitive levels: Remember, Understand, Apply, Analyze, Evaluate and Create. Two researchers both with expertise in educational assessment and taxonomy-based categorization, independently assigned taxonomy levels to each MCQ. The Inter-Rater Reliability (using Cohen's Kappa) is 0.67. For cases where their classifications differed (N = 309), a third researcher who was also an expert in educational assessment reviewed the discrepancies and selected the more accurate classification. The final distribution of categorized MCQs is presented in Table 1.

2.3 Distractor Generation

To generate distractors for MCQs, this study employed GPT-4 API, a state-of-the-art large language model that have been frequently used in recent studies to automate the creation of distractors or answers (Doughty et al., 2024; Rodrigues et al., 2024). To ensure consistency and relevance, a standardized prompt structure was designed, incorporating four components (i.e., the question stem, the correct answer, the corresponding Bloom's taxonomy level, and specific instructional constraints). The inclusion of Bloom's taxonomy categories is to guide GPT-4 in generating distractors that match the intended cognitive level of each question. The prompt explicitly instructed GPT-4 to generate three incorrect but plausible distractors in a structured format. The primary objective was to ensure that the generated distractors maintained semantic and cognitive alignment with the question stem and correct answer while varying in difficulty levels according to Bloom's taxonomy. GPT-4 was accessed through OpenAI's API with hyperparameters set to optimize creativity and reliability, specifically using a temperature of 0.7 and top-p of 0.9 to balance diversity and control in response generation.

Following the generation process, all distractors underwent a two-step validation process to ensure quality and reliability. First, an automated filtering step was applied to identify and eliminate instances of empty or erroneous generations. In cases where the generated output was incomplete or contained errors, GPT-4 was re-invoked to regenerate the distractors. Second, a human expert review was conducted to refine the dataset. Any extraneous content beyond the intended distractors was identified and removed. The validated distractors were then incorporated into the final MCQ dataset for further analysis. Distractors that duplicate teacher-created distractors are included.

2.4 Human-AI Collaboration Distractor Evaluation

To evaluate the quality of AI-generated distractors in comparison to teacher-created distractors, we employed a human-AI collaborative framework involving both human experts and state-of-the-art generative AI models (i.e., DeepSeek-V3 (DeepSeek-AI et al., 2024) and GPT-4o (OpenAI et al., 2024)). This hybrid evaluation approach was designed to ensure a unbiased assessment of distractor plausibility and pedagogical effectiveness while optimizing human resource allocation. The evaluation process consisted of three key steps. First, for each MCQ, a total of six distractors (three generated by teachers and three by GPT-4) were randomly shuffled and anonymized to prevent evaluators from identifying their source. Then, each human expert and GenAI model independently assessed all six distractors and selected the three most plausible and pedagogically effective options. The GenAI models were instructed using the following prompt: "You are a programming/statistics course instructor

evaluating distractors for multi-choice questions. Select the three most effective distractors based on plausibility, similarity to the correct answer, and ability to challenge the test-taker." After collecting the selections, the final set of three distractors was determined through a ranking process. The distractor with the highest selection frequency across all evaluators was chosen first (e.g., Distractor 4 in Figure 2). Among the remaining distractors with the same selection frequency, those endorsed by both AI and human evaluators were prioritized (e.g., Distractor 2 and Distractor 3). This selection process ensured that the final distractors represented the most effective choices from human and AI.

To determine whether GPT-4-generated distractors were of comparable or superior quality to teacher-created ones, we aggregated the selection results across all evaluators. The frequency with which each distractor was selected served as an indicator of its quality. If AI-generated distractors were chosen more frequently than teacher-created distractors across multiple MCQs, it suggested that AI could generate plausible and effective distractors. Conversely, if teacher-created distractors were selected more often, it indicated that human-authored distractors remained more pedagogically valid. Finally, a Wilcoxon Signed-Rank Test was conducted at the six Bloom's taxonomy levels to assess whether the differences in selection rates between AI-generated and teacher-created distractors were significant.

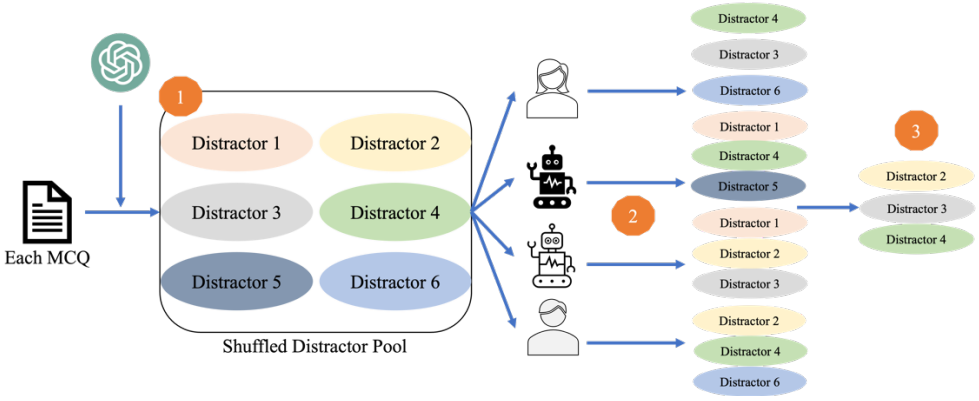


Figure 2 The Human-AI collaboration evaluation process.

3. Results

3.1 AI-Generated Distractors

A descriptive analysis of the GPT-4-generated distractors is shown in Table 2. Length represents the number of words in each distractor. Flesch Reading Ease and Flesch-Kincaid Grade Level measure readability, and the results indicate similar readability levels across both courses. Distractor Similarity was calculated using the Jaccard similarity between the three generated distractors for each question. For Answer Similarity, the Levenshtein distance between each distractor and the correct answer was computed, where a larger distance indicates greater dissimilarity. Due to the presence of numeric distractors, the minimum Levenshtein distance across both courses ranged from 2 to 3, while the average distance was between 40 and 68, indicating a reasonable level of distinction from the correct answers.

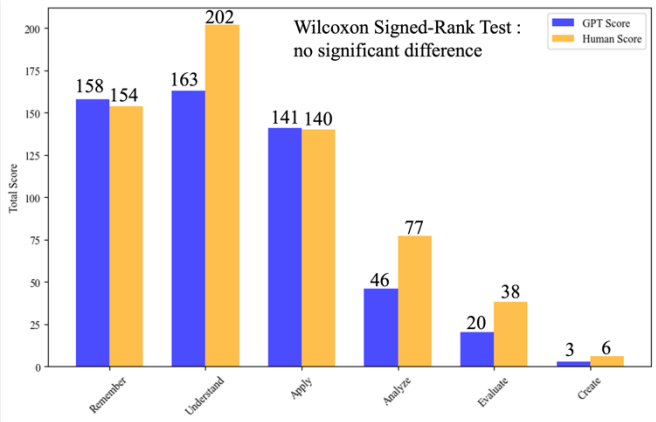
Table 2 Description of the AI-generated distractors

	Programming			Statistic		
Metric	Min	Mean (SD)	Max	Min	Mean (SD)	Max
Length	3	33.9 (26.2)	95	3	52.2 (50.7)	352
Reading	31.3	68.4 (38.2)	121.2	29.2	78.9 (34.3)	121.2
FKGL	3.5	5.1 (5.6)	16.5	3.5	4.3 (5.7)	19.2
J Similarity	-	0.01 (0.06)	1	-	0.01 (0.08)	1
L Similarity	3	41.0 (27.2)	210	2	67.4 (54.7)	351

Reading: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level, J Similarity: Flesch-Kincaid Grade Level, L Similarity: Answer Similarity (Levenshtein)

3.2 AI-Generated vs Human-created Distractors

From Figure 3, it is evident that for programming courses, ChatGPT and AI perform consistently on MCQs related to Remember, Apply, and Create levels of Bloom's taxonomy. However, for other types such as Understand, Analyze, and Evaluate, human distractors score higher. Further analysis using the Wilcoxon signed-rank test reveals no significant differences in scores between human and AI for these types of MCQs. The table indicates that there are nine questions each where human and AI scores are equal for Remember and Understand questions. For statistics courses, AI and human teachers perform similarly on Remember, Apply, and Create questions. However, human teachers perform better in the other three dimensions. Further Wilcoxon analysis shows significant differences in Understand, Analyze, and Evaluate, with the following results: ($p = 0.002$, with a small effect size = 0.17), ($p < 0.001$, with a large effect size = 0.58), and ($p = 0.008$, with a medium effect size = 0.32). This indicates that AI has a noticeable gap in generating distractors for Analyze-type MCQs compared to human teachers.



Bloom Level	Count Equal	Count GPT_Higher	Count GPT_Lower
Analyze	1	11	28
Apply	4	39	44
Create	0	1	2
Evaluate	0	5	13
Remember	9	42	40
Understand	9	41	61

Figure 3 The evaluation results

4. Discussion

Advancements in AI-driven distractor generation offer promising opportunities for automating MCQ creation. Our findings highlight both the strengths and limitations of AI-generated distractors compared to human-crafted ones. GPT-4-generated distractors performed comparably to human-created ones in lower-order cognitive skills (Remember, Apply, and Create), aligning with prior research on pre-trained language models' ability to generate diverse, context-aware distractors (Bulathwela et al., 2023; Lelkes et al., 2021). Low distractor similarity scores further suggest AI-generated options are varied and non-redundant. However, challenges arise in higher-order cognitive skills (Understand, Analyze, Evaluate). While no significant differences emerged in programming MCQs, which aligns with previous work focus on programming distractors (Doughty et al., 2024), statistics MCQs showed substantial gaps in Understand, Analyze, and Evaluate MCQs. The large effect size in Analyze suggests AI struggles with nuanced reasoning, consistent with prior studies on the difficulty of generating pedagogically effective distractors for complex tasks (Bitew et al., 2023; Wei et al., 2022). This limitation likely stems from AI's challenges in semantic reasoning and conceptual understanding. Though strategies like few-shot and chain-of-thought prompting aim to improve distractor plausibility (McNichols et al., 2023), our results indicate their shortcomings in abstract reasoning-heavy disciplines like statistics. Building on previous studies (Atchley et al., 2024; Doughty et al., 2024), this study proposes the human-AI collaborative assessment of distractor quality, which to some extent alleviates the workload of fully manual evaluation. Future research should refine hybrid methods integrating fine-tuning and prompting, develop

more robust evaluation metrics, and explore human-in-the-loop validation. Additionally, instructional MCQs, which demand multi-modal analysis, warrant further investigation. By improving AI-assisted question design, this research supports scalable, effective assessment tools for educators.

Acknowledgements

The project is funded by the National Science Foundation (NSF) of the United States under grant numbers 1503196 and 2105695.

References

Atchley, P., Pannell, H., Wofford, K., Hopkins, M., & Atchley, R. A. (2024). Human and AI collaboration in the higher education environment: Opportunities and concerns. *Cognitive Research: Principles and Implications, 9*(1), 20. https://doi.org/10.1186/s41235-024-00547-9

Bitew, S. K., Deleu, J., Develder, C., & 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* (pp. 48–63). Springer. https://doi.org/10.1007/978-3-031-74627-7_4

Bulathwela, S., Muse, H., & Yilmaz, E. (2023). Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education* (pp. 327–339). Springer. https://doi.org/10.1007/978-3-031-36272-9_27

Chung, H.-L., Chan, Y.-H., & Fan, Y.-C. (2020). A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4390–4400). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.393

DeepSeek-AI, Liu, A., Feng, B., Xue, B., et al. (2024). Deepseek-v3 technical report. *arXiv*. https://arxiv.org/abs/2412.19437

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024, January 29–February 2). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In *Proceedings of the Australian Computing Education Conference (ACE 2024)* (p. 10). ACM. https://doi.org/10.1145/3636243.3636256

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine Translation, 23*, 105–115. https://doi.org/10.1007/s10590-009-9059-4

Lelkes, A. D., Tran, V. Q., & Yu, C. (2021). Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021* (pp. 2501–2511). https://doi.org/10.1145/3442381.3449892

Lin, C. Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics.

McNichols, H., Feng, W., Lee, J., Scarlatos, A., Smith, D., Woodhead, S., & Lan, A. S. (2023). Exploring automated distractor and feedback generation for math multiple-choice questions via in-context learning. *CoRR*. arXiv:2308.03234. https://doi.org/10.48550/arXiv.2308.03234

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., et al. (2024). GPT-4 technical report. *arXiv*. https://arxiv.org/abs/2303.08774

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Rodrigues, L., Pereira, F. D., Cabral, L., Gašević, D., Ramalho, G., & Mello, R. F. (2024). Assessing the quality of automatic-generated short answers using GPT-4. *Computers and Education: Artificial Intelligence, 7*, 100248. https://doi.org/10.1016/j.caeai.2024.100248

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://doi.org/10.5555/3600270.3602070

Zhou, X., Luo, S., & Wu, Y. (2020). Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 9725–9732. https://doi.org/10.1609/aaai.v34i05.6522