LLM-Based Approaches for Detecting Gaming the System in Self-Explanation

Jiayi Zhang
University of Pennsylvania
Philadelphia, PA, United States
joycez@upenn.edu

Ryan S. Baker
Adelaide University
Adelaide, Australia
ryanshaunbaker@gmail.com

Bruce M. McLaren
Carnegie Mellon University
Pittsburgh, PA, United States
bmclaren@andrew.cmu.edu

Abstract

Self-explanation supports deeper learning by prompting students to articulate their reasoning and connect new concepts with prior knowledge. Open-ended self-explanation questions promote elaborative processing and help address knowledge gaps. However, these benefits may be undermined when students game the system — a maladaptive learning strategy where students exploit the learning environment rather than engaging in meaningful learning. While previous studies have successfully detected this behavior in students' interactions with learning activities, this study focuses on identifying such behavior in students' open-ended responses within a math digital learning game. We evaluated two large language model (LLM)-based approaches: one using sentence embeddings and another using a prompt-based method. Both showed acceptable performance, but the embedding-based model outperformed the prompt-based one. Error analysis revealed the prompt-based model struggled with short, low-context responses and produced false positives when students referenced using hints. Consistent with earlier findings, we showed that higher rates of gaming behavior in open-ended responses negatively correlated with learning gains.

1 Introduction

Self-explanation, an important pedagogical strategy, has been frequently used in classrooms to facilitate learning. During this process, students articulate their reasoning, connect new information with prior knowledge, and identify gaps in their understanding (Fonseca and Chi, 2011; Wylie and Chi, 2014). Self-explanation can be self-initiated or externally prompted. Previous studies have shown that self-explanation leads to improved performance, deeper conceptual understanding, and better long-term retention (Bisra et al., 2018; VanLehn et al., 1992). In mathematics learning, students who engage in self-explanation are more likely to develop a more robust understanding of problems and improve their ability to transfer knowledge to novel situations (McEldoon et al., 2013; Rittle-Johnson, 2006).

Given these benefits, self-explanation questions have been increasingly integrated into digital learning platforms. However, due to the limitations of digital learning systems—which, until recently, had a limited ability to process natural language and provide feedback—self-explanation questions have often been designed in a closed-ended format, such as multiple-choice, fill-in-the-blank questions, or sentence builders (McLaren et al., 2022). Nonetheless, open-ended self-explanation questions "may invite elaborative processing better adapted to each learner's unique gaps in knowledge" (Bisra et al., 2018) and encourage deeper cognitive processing (Kwon et al., 2011). A recent study comparing three self-explanation formats (multiple-choice, fill-in-the-blank, and open-ended) found that students who answered open-ended self-explanation questions achieved the greatest learning gains (McLaren et al., 2022).

However, failing to engage meaningfully with these self-explanation questions can potentially diminish the positive effects. In gaming the system, a disengaged behavior and maladaptive learning strategy, students attempt to succeed by exploiting system properties rather than engaging in meaningful learning, resorting to behaviors such as systematic guessing or abusing hints (Baker et al., 2008). Gaming the system has been observed across platforms and is consistently associated with lower learning gains and long-term negative outcomes (e.g., Baker et al. (2006b); Cocea et al. (2009)). In a previous study, the negative effects of gaming have also been demonstrated within (non-open-ended) self-explanation questions, in which students who had a higher rate of gaming were associated with lower learning gain. Furthermore, the rate of gaming in the self-explanation moderated the differences

in learning between boys and girls (Baker et al., 2024).

As such, to support interventions, gaming detectors have been developed in the past to identify instances when students game the system (Li et al., 2022; Xia et al., 2020). However, most of these detectors are designed for close-ended questions, which identify gaming based on interaction patterns with learning activities. A few gaming detectors for text-based open-ended responses have primarily focused on response patterns (e.g., detecting repetition in open-ended responses) rather than analyzing the semantic content of the inputs (Darvishi et al., 2022). For example, identifying instances where students game the system by cycling through answers, entering responses such as "It will be 7.1", "It will be 7.2", "It will be 7.3". As a result, a significant gap remains in detecting gaming behaviors in the open-ended responses.

The advancement of large language models (LLMs) presents an opportunity for this use case. Trained on vast amounts of text data, these models have demonstrated capabilities in processing, understanding, and generating natural language with high accuracy (Brown et al., 2020). As a result, LLMs have been increasingly used to analyze and categorize textual data, presenting an opportunity to perform classification tasks such as assessing the correctness or relevance of self-explanations (Nguyen et al., 2023) or identifying the presence or absence of gaming in open-ended responses. One common approach to leveraging LLMs for classification tasks is through sentence embeddings, where text inputs are transformed into high-dimensional vectors that capture semantic meaning. These embeddings can then be input into machine learning models to categorize responses. Alternatively, prompt-based methods (e.g. Generative Pre-trained Transformer; GPT) frame classification tasks as text-generation problems, allowing pre-trained LLMs to infer labels based on contextual prompts. Several studies have found that classifying embeddings outperforms prompt-based approaches in various classification tasks (Liu et al., in press; Hutt et al., 2024). Recent studies have explored prompt engineering, examining how one-shot (providing one example), few-shot (providing a few examples), adding context (Xiao et al., 2023), modifying prompt structure (White et al., 2023), and defining roles influence model performance (Hou et al., 2024). However, less research has explored where the two approaches diverge and under what conditions or context one approach is more effective than the other, evaluating and comparing the validity and reliability of the two approaches for classification tasks.

In this study, we explored the use of large language models (LLMs) to detect gaming the system in open-ended responses to self-explanation questions within a math digital learning game. We identified gaming behavior using both an embedding-based and a prompt-based approach and compared their performance. To understand where the two approaches diverge, we conducted an error analysis examining the types of errors each approach is prone to, highlighting the context under which one approach might be more efficient than the other. Lastly, we applied the best-performing model to the full dataset and conducted analyses to examine the relationships between gaming during the self-explanation step and learning gains within this learning system. By detecting gaming the system in this additional context, we enhance our understanding of how broadly this phenomenon occurs and enable learning technologies to intervene in a wider range of contexts. Additionally, the comparison between the two approaches contributes to the growing body of research on leveraging LLMs for text classification.

2 Methods

2.1 Learning Platform and Data

Student log data were collected from Decimal Point, a single-player web game designed to motivate middle-school students to learn decimal concepts (McLaren, 2024; McLaren et al., 2017). Students wander through a virtual amusement park and play a variety of mini-games that incorporate decimal challenges, such as sorting decimals. In the version of the game where the data was collected, students were first asked to solve a problem (problem-solving step) and then prompted to reflect on how they solve the problem and explain their reasoning with an open-ended self-explanation question (self-explanation step) (McLaren et al., 2022). To assure that students expend at least minimal effort in answering the self-explanation questions, the response needed to contain at least four words with at least one of the words from a relevant list (including common misspellings) that would legitimately be found in a correct explanation. Students could make multiple attempts and could only move to the next question once the response meets these

criteria.

To investigate LLM's ability at detecting gaming in open-ended responses, we collected the text-based responses submitted by 212 students and delineated them into clips, with each clip containing all the attempts (responses) a student submitted at answering a self-explanation question. In total, 2553 clips were extracted. We also collected students' pre-test, post-test, and delayed post-test scores.

2.2 Coding Gaming the System

Text replay coding was conducted to establish ground truth. In text replays, human coders examine each clip and determine the presence or absence of gaming the system using a codebook (Baker et al., 2006a). The codebook was developed through an iterative process to ensure that the behaviors classified as gaming aligned with previous conceptualizations (e.g. as defined in Baker et al. (2008)) and were salient in the dataset. Through this process, we developed a codebook consisting of three criteria: (1) a low degree of semantic difference between consecutive responses – e.g. changing between highly related alternatives, (2) systematically cycling through modifications to responses or potential multiple answers, and (3) making a conceptual or functional change between responses (e.g., identifying a concept versus suggesting an action, trying to figure out what category of response is needed without thinking through the question) in conjunction with the previous two criteria. The gaming criteria and examples are presented in Table 1.

Using the codebook, two coders first independently coded the same set of data to establish inter-rater reliability (κ = 0.8). Once consensus was reached, the coders proceeded to code a total of 1,465 clips from 116 students, of which 8.9% were positive (gaming) clips.

2.3 Approach 1: Detecting Gaming with Sentence Embeddings

To train models that automatically detect gaming, we first con-catenated textual responses from all attempts within a clip, separating each attempt with a period. We then vectorized the text using two sentence embedding models: the Universal Sentence Encoder Large v5 (USE) developed by Google, which generates a 512-dimensional vector for each entry (Cer et al., 2018), and sentence-embedding-3-short developed by OpenAI, which produces a 1,536-dimensional vector (Neelakantan et al., 2022).

For each set of embeddings, we trained a neural network model with one hidden layer to predict the presence or absence of gaming. The models were evaluated using 5-fold student-level cross-validation. Model performance was evaluated using the average Area Under the Receiver Operating Characteristic Curve (AUC) and Kappa.

2.4 Approach 2: Detecting Gaming using Prompt-Based Model

For prompt-based methods, we leveraged both zero-shot and one-shot prompting techniques, providing the GPT-4-turbo model with the definition of gaming the system and the three criteria from the codebook for zero-shot prompting, and the corresponding examples (as listed in the codebook) for one-shot prompting. The exact prompt used for zero-shot prompting is presented below. For one-shot prompting, examples were added to the prompt. The temperature was set to 0 to minimize randomness. To account for the stochastic nature of GPT, we ran the prompt three times to assess consistency across iterations. The final prediction was determined using majority voting across the three outputs. The predictions were evaluated against the ground truth using AUC and Kappa.

> "Review the provided text and code it based on the construct: gaming the system. The definition of this construct is: a maladaptive learning strategy where students attempt to succeed by exploiting properties of a learning environment. Some criteria of gaming the system in open-ended responses include: 1) a low degree of semantic difference between responses, 2) cycling through multiple answers/ modifications to their responses, or 3) conceptual or functional change between responses (e.g., identifying a concept versus suggesting an action) accompanied by the previous two criteria. After reviewing the text, assign a code of '1' if you believe the text exemplifies gaming the system, or a '0' if it does not. Your response should only be '1' or '0'. TEXT TO BE REVIEWED: [TEXT]"

Gaming Criteria	Attempt 1	Attempt 2	Attempt 3	Attempt 4
Minor Semantic Difference	I need to move it vertically	Move side to side	–	–
Cycling through Modifications	It will be 7.1	It will be 7.2	It will be 7.3	–
Conceptual or Functional Change	It is 1.7	It is 1.9	By adding	By subtracting

Table 1: Examples of gaming behaviors across multiple attempts.

3 Results

3.1 Model Performance

As shown in Table 2, with 5-fold student-level cross-validation, the neural network model built using sentence embeddings from the Universal Sentence Encoder achieved an average AUC of 0.902 and a Kappa of 0.535. The neural network model using sentence-embedding-3-short as the encoder performed better, reaching an average AUC of 0.935 and a Kappa of 0.564. In contrast, the prompt-based model with zero-shot prompting achieved an AUC of 0.699 and a Kappa of 0.345, and an AUC of 0.754 and a Kappa of 0.358 with one-shot prompting. We also recorded the number of false positive and false negative cases for each model, which is discussed in the next section.

3.2 Error Analysis

To examine differences in prediction accuracy across the models, we conducted an error analysis using both quantitative and qualitative methods, counting the number of type I and type II errors as well as reviewing the responses the models misclassified.

As shown in Table 2, both sentence embedding models were more likely to make Type II errors (false negatives) than Type I errors (false positives), meaning they incorrectly assessed the student as not gaming when the response actually demonstrated gaming behaviors. In contrast, the prompt-based models were more prone to Type I errors (false positives) than Type II errors (false negatives), predicting gaming when the student was not actually gaming. Additionally, Type I errors were twice as frequent for the prompt-based models compared to the sentence embedding models.

To better understand where the models failed to make accurate predictions, we examined the misclassified cases, analyzing responses in which there was a discrepancy between the sentence embedding approach and the prompt-based approach. Of the 1,465 responses, 178 were correctly classified by both sentence embedding models (Universal Sentence Encoder and sentence-embedding-3-short) but misclassified by at least one of the prompt-based approaches. Among these, 25 had a true label of gaming, and 153 had a true label of not gaming.

Upon examining these cases, we identified several patterns. One common pattern among **false positives** for the prompt-based models was responses that are not considered gaming in this particular dataset but could be considered gaming if gaming were defined more broadly. For example, some responses mentioned the use of hints. In one instance where the prompt-based model falsely classified the behavior as gaming, the student (somewhat oddly) stated "Always remember to use the hint button. It gives you the answer if you click until it doesn't say 'next,' and you should get the answer correct if you follow what it says." Another false positive example is when a student said, "22.0. You have to add. 22.0. You can look at the hints to find the answer. You can find this answer by adding 17.6 + 4.4." In these cases, the prompt-based model flagged the responses as gaming likely due to mentions of hints, but they may not strictly align with the definition of gaming behavior in this specific context.

Comparing between zero-shot and one-shot prompting for the false positive cases, we noticed that the majority of cases misclassified by the one-shot model were responses that repeated themselves without any semantic changes. For example, when asked, "Is 0.2 bigger or smaller than 0.22? How do you know?", a student responded, "It is smaller. It is smaller." The model with one-shot prompting misclassified this as gaming, whereas the zero-shot model correctly classified it as not gaming, as there was no semantic difference between the two entries, and it didn't imply a cycling behavior.

A common pattern among **false negatives** for the prompt-based model with zero-shot prompting was that shorter responses lacked sufficient con-

Model	AUC (stdev)	Kappa (stdev)	False Positive	False Negative
Universal sentence encoder	0.902 (0.038)	0.535 (0.087)	52	73
sentence-embedding-3-short	0.935 (0.026)	0.564 (0.088)	44	70
Prompt-based zero-shot	0.699	0.345	112	68
Prompt-based one-shot	0.754	0.358	169	48

Table 2: Classification results and total errors.

text for the model to accurately interpret gaming behavior. For example, when asked the same question, "Is 0.2 bigger or smaller than 0.22? How do you know?" a student responded, "Smaller. Bigger. Bigger. Bigger. Smaller. Smaller because 0.22 has an extra digit than 0.2." Due to the brevity of the response, the model may have struggled to contextualize it properly, leading to a misclassification. However, this is less frequent with one-shot prompting, possibly because of the brevity in the examples provided.

Altogether, these patterns suggest that the prompt-based model may struggle with nuanced cases where gaming behaviors depend on context, leading to predictions that are not context-specific. Specifically, it tends to misclassify responses that mention hints or shortcuts as gaming, even when they might not strictly fit the definition based on the current operationalization. Compared to zero-shot, one-shot prompting is also more prone to Type II errors, misclassifying cases where students repeat responses as gaming rather than as recycling responses with minimal semantic changes. Conversely, prompt-based approach struggles to detect gaming in shorter responses that lack sufficient context, especially when examples are not provided.

The same qualitative approach was conducted to evaluate the predictions of the embedding-based models, focusing on responses that were correctly classified by both prompt-based models but misclassified by at least one of the embedding-based models. Of the 1,465 responses, 84 were correctly classified by both prompt-based models (zero-shot and one-shot) but misclassified by at least one of the embedding-based models. Among these, 32 had a true label of gaming, and 52 had a true label of not gaming.

By analyzing the **false negative** cases, we found that, similar to zero-shot prompting, sentence-embedding models are prone to Type II errors when responses are brief and seemingly disjointed. This issue is especially apparent when key explanatory words (such as "because") are missing. For example, when asked, "Is 0.456 to the left of 0 or to the right of 0 on the number line? How do you know?", one student responded, "Right. 0.5. Left. 0.45. 0.45 to the right." Another example comes from the question, "Is 6.5 bigger or smaller than 6.41? How do you know?", to which a student responded, "6.5 is smaller. 6.41 is smaller. 6.41 is bigger." These responses clearly reflect cycling behavior, even though they lack explanatory words (such as "because") that directly address the question's explanatory prompt. Sentence-embedding models failed to detect gaming in such cases possibly because they rely on overall semantic similarity to the example cases (e.g., frequent usage of explanatory terms) and lack the contextual understanding needed to recognize patterns like repetitive guessing or cycling.

3.3 Gaming the System and Learning Gains

After applying the best model (the model trained using embeddings derived from sentence-embedding-3-short) to the full dataset (2,553 clips), we found that students' detected frequency of gaming was negatively correlated with the pre-test ($r = -0.233$, $p = 0.058$), post-test ($r = -0.312$, $p = 0.010$), and delayed post-test ($r = -0.355$, $p = 0.003$). We found that gaming frequency was not correlated with normalized learning gains between the pre-test and post-test ($r = -0.121$, $p = 0.329$), but was negatively and significantly correlated with normalized learning gains between the pre-test and delayed post-test ($r = -0.247$, $p = 0.044$).

4 Discussion and Conclusion

4.1 Main Findings

Self-explanation promotes deeper learning by helping students articulate their reasoning and connect new information with prior knowledge. Open-ended self-explanation questions, in particular, foster more elaborative processing, allowing students to address their unique knowledge gaps. However,

these benefits can be undermined when students disengage and attempt to game the system. This study addresses this challenge by introducing an automated approach to detect gaming in open-ended responses using large language models (LLMs). Specifically, we compare a sentence embedding-based method with a prompt-based approach. By identifying gaming behavior in real time, this method can support targeted interventions, such as adaptive feedback, to help students re-engage and maximize the benefits of self-explanation.

Our results show that while all models demonstrate reliable performance in detecting gaming in open-ended responses, the sentence embedding-based approach, particularly the OpenAI sentence-embedding-3-short model, outperformed the prompt-based method, achieving an AUC of 0.935 and a Kappa of 0.564. While the prompt-based model was easier to implement, it was more prone to false positives, frequently misclassifying responses that mentioned hints or repeated responses as gaming. These results highlight the challenges of using prompt-based models for nuanced classification tasks, particularly when the definition of the target behavior is context-dependent

Additionally, both prompt-based and sentence-embedding-based models struggled with shorter, context-poor responses, leading to false negatives. However, this issue can be attenuated with one-shot prompting.

Overall, the comparison between the two approaches suggests that sentence-embedding is more conservative in detecting gaming, making it more prone to Type II than Type I errors, at least for this application. On the other hand, the prompt-based approach—possibly due to access to additional contextual information provided in the prompt is more liberal and less context-specific, making it more prone to Type I than Type II errors. These findings may suggest a direction for future study to explore the possibility of combining the two approaches and leveraging them for their strengths. It is also possible to adapt the model selection based on the data as well as the desired outcomes.

Furthermore, we found that the frequency of detected gaming behavior was negatively correlated with students' pre-test, post-test, delayed post-test scores, and delayed learning gains, suggesting that gaming the system in this context is also associated with lower learning outcomes. This aligns with previous research that has consistently linked gaming behavior with reduced learning gains (Baker et al., 2008; Cocea et al., 2009).

4.2 Future Work

We acknowledge the following limitations. First, it is possible that the prompt-based model's performance may have been constrained by the limited prompt engineering employed in this study, for instance, not providing more specific context information for self-explanations. Future work could explore more sophisticated prompting strategies, such as few-shot learning, where the model is provided with more than one labeled example to improve its performance. Additionally, fine-tuning the LLM on a domain-specific dataset could further enhance its ability to detect gaming especially in contexts where nuanced semantic understanding is critical.

Second, the generalizability of our findings may be limited by the specific context in which gaming is being operationalized. Future studies should validate these approaches in other learning environments and with more diverse datasets. This would help determine whether the observed patterns hold across different digital learning contexts and student populations.

Finally, while our study focused on detecting gaming behavior, future research could explore the possibility of distinguishing specific gaming behaviors (e.g., minor semantic differences or cycling through modifications) and examine whether they impact learning outcomes differentially.

4.3 Conclusion

In contrast to previous gaming detectors based on interaction data, this study demonstrates the potential of using LLMs to detect gaming behavior in open-ended self-explanation responses by identifying gaming based on the semantic meaning of text-based responses. Our findings suggest that sentence embedding-based approaches are more effective than prompt-based methods for this task, possibly because the definition of gaming the system is context-dependent. Consistent with prior research, we found that gaming in open-ended self-explanation questions is also negatively correlated with learning gains, emphasizing its detrimental impact and the need for intervention. The ability to detect gaming in open-ended responses opens new possibilities for intervention and support in digital learning environments, helping ensure that students engage meaningfully with self-explanation tasks and achieve better learning outcomes.

References

R.S. Baker, A.T. Corbett, and A.Z. Wagner. 2006a. Human classification of low-fidelity replays of student actions. In *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, pages 29–36.

R.S. Baker, J.E. Richey, J. Zhang, S. Karumbaiah, J.M. Andres-Bray, H.A. Nguyen, J.M.A.L. Andres, and B.M. McLaren. 2024. Gaming the system mediates the relationship between gender and learning outcomes in a digital learning game. *Instructional Science*.

R.S.J. d. Baker, A.T. Corbett, K.R. Koedinger, S. Evenson, I. Roll, A.Z. Wagner, M. Naim, J. Raspat, D.J. Baker, and J.E. Beck. 2006b. Adapting to when students game an intelligent tutoring system. In *Intelligent Tutoring Systems*, pages 392–401. Springer Berlin Heidelberg.

R.S.J.D. Baker, A.T. Corbett, I. Roll, and K.R. Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18(3):287–314.

K. Bisra, Q. Liu, J.C. Nesbit, F. Salimi, and P.H. Winne. 2018. Inducing self-explanation: a meta-analysis. *Educational Psychology Review*, 30(3):703–725.

T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, and T. Henighan. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R.S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. 2018. Universal sentence encoder. *arXiv preprint*, arXiv:1803.11175.

Mihaela Cocea, Arnon Hershkovitz, and Ryan S.J.d. Baker. 2009. The impact of off-task and gaming behaviors on learning: Immediate or aggregate? In *Frontiers in Artificial Intelligence and Applications*, pages 207–514. IOS Press.

A. Darvishi, H. Khosravi, S. Sadiq, and D. Gašević. 2022. Incorporating ai and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*, 53(4):844–875.

B.A. Fonseca and M.T. Chi. 2011. Instruction based on self-explanation. In *Handbook of Research on Learning and Instruction*, pages 310–335. Routledge.

C. Hou, G. Zhu, J. Zheng, L. Zhang, X. Huang, T. Zhong, S. Li, H. Du, and C.L. Ker. 2024. Prompt-based and fine-tuned gpt models for context-dependent and -independent deductive coding in social annotation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 518–528, Kyoto, Japan.

S. Hutt, A. DePiro, J. Wang, S. Rhodes, R.S. Baker, G. Hieb, S. Sethuraman, J. Ocumpaugh, and C. Mills. 2024. Feedback on feedback: Comparing classic natural language processing and generative ai to evaluate peer feedback. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 55–65, Kyoto, Japan.

K. Kwon, C.D. Kumalasari, and J.L. Howland. 2011. Self-explanation prompts on problem-solving performance in an interactive learning environment. *Journal of Interactive Online Learning*, 10(2).

Y. Li, X. Zou, Z. Ma, and R.S. Baker. 2022. A multi-pronged redesign to reduce gaming the system. In *International Conference on Artificial Intelligence in Education*, pages 334–337.

X. Liu, A.F. Zambrano, R.S. Baker, A. Barany, J. Ocumpaugh, J. Zhang, M. Pankiewicz, N. Nasiar, and Z. Wei. in press. Qualitative coding with gpt-4: Where it works better. *Journal of Learning Analytics*.

K.L. McEldoon, K.L. Durkin, and B. Rittle-Johnson. 2013. Is self-explanation worth the time? a comparison to additional practice. *British Journal of Educational Psychology*, 83(4):615–632.

B.M. McLaren. 2024. Decimal point: A decade of learning science findings with a digital learning game. In P. Ilic, I. Casebourne, and R. Wegerif, editors, *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*, pages 145–203. Springer Nature Switzerland.

B.M. McLaren, D.M. Adams, R.E. Mayer, and J. Forlizzi. 2017. A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning*, 7(1):36–56.

B.M. McLaren, J.E. Richey, H.A. Nguyen, and M. Mogessie. 2022. Focused self-explanations lead to the best learning outcomes in a digital learning game. In *Proceedings of the 16th International Conference on Learning Science*, pages 36–56.

A. Neelakantan and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint*, arXiv:2201.10005.

H.A. Nguyen, H. Stec, X. Hou, S. Di, and B.M. McLaren. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*, pages 278–293, Cham.

B. Rittle-Johnson. 2006. Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, 77(1):1–15.

K. VanLehn, R.M. Jones, and M.T.H. Chi. 1992. A model of the self-explanation effect. *Journal of the Learning Sciences*, 2(1):1–59.

J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D.C. Schmidt. 2023. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint*, arXiv:2302.03459.

R. Wylie and M.T. Chi. 2014. The self-explanation principle in multimedia learning. In *The Cambridge Handbook of Multimedia Learning*, pages 413–432.

M. Xia, Y. Asano, J.J. Williams, H. Qu, and X. Ma. 2020. Using information visualization to promote students' reflection on "gaming the system" in online learning. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, pages 37–49, Virtual Event USA.

Z. Xiao, X. Yuan, Q.V. Liao, R. Abdelghani, and P.-Y. Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *28th International Conference on Intelligent User Interfaces*, pages 75–78, Sydney, NSW, Australia.