Beyond the Hint: Using Self-Critique to Constrain LLM Feedback in Conversation-Based Assessment

Tyler Burleigh
Khan Academy
tylerb@khanacademy.org

Jenny Han
Khan Academy
jennyhan@khanacademy.org

Kristen DiCerbo
Khan Academy
kristen@khanacademy.org

Abstract

Large Language Models in Conversation-Based Assessment tend to provide inappropriate hints that compromise validity. We demonstrate that self-critique – a simple prompt engineering technique – effectively constrains this behavior. Through two studies using synthetic conversations and real-world high school math pilot data, self-critique reduced inappropriate hints by 90.7% and 24-75% respectively. Human experts validated ground truth labels while LLM judges enabled scale. This immediately deployable solution addresses the critical tension in intermediate-stakes assessment: maintaining student engagement while ensuring fair comparisons. Our findings show prompt engineering can meaningfully safeguard assessment integrity without model fine-tuning.

1 Background

1.1 Introduction

Conversation-Based Assessment (CBA) represents an innovative approach to educational evaluation. In CBA, students engage in dialogue with a chatbot while being assessed, which can improve test score validity (Yildirim-Erbasli and Bulut, 2023). Unlike traditional formats, CBA enables natural language responses that expand construct coverage (Bejar, 2017) while providing two unique assessment advantages: immediate, tailored feedback to enhance engagement, and follow-up questions that probe deeper understanding when initial responses are incomplete.

While CBA has shown promise in low-stakes formative assessments, intermediate-stakes assessments present a unique challenge (Perie et al., 2009). These assessments require both student engagement to ensure validity (Eklöf, 2010; Finn, 2015) and standardized conditions to enable fair comparisons between students. This creates tension between providing motivating feedback and maintaining assessment standardization.

The integration of Large Language Models (LLMs) into CBA systems presents both an opportunity and a challenge. LLMs excel at providing supportive, encouraging responses that could enhance student engagement – a critical factor for assessment validity. They achieve this through training that maximizes human preferences (Ziegler et al., 2020). However, this same preference-maximizing behavior leads LLMs to naturally provide overly helpful responses. These responses may include inappropriate hints, solutions, or answers (Jones and Bergen, 2024). For assessments where protecting validity and comparability is critical, LLM behavior must be carefully constrained to harness engagement benefits while preventing inappropriate assistance (Puech et al., 2025).

1.2 Constraining LLM behavior

The critical need to prevent inappropriate assistance in assessment contexts makes methods for constraining LLM behavior essential. While model tuning can modify behavior through weight updates, prompt engineering (PE) offers a more accessible approach using carefully crafted instructions and code-based techniques (Vijayan and Vengathattil, 2025).

Among PE techniques for behavioral constraint (Sahoo et al., 2024), self-critique shows particular promise. This technique uses the LLM to critique and revise its own responses (He et al., 2025), demonstrating effectiveness at reducing hallucinations (Dhuliawala et al., 2023) and performing well as a self-critic for short inputs (He et al., 2025), making it well-suited for assessment applications where responses are typically brief.

1.3 Evaluation methodology

Rigorous measurement is essential for evaluating LLM behavior in assessment contexts. Evaluating whether an LLM gives inappropriate hints requires measurement methodology borrowed from

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 79–85
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

social science (Ameli et al., 2024; Wallach et al., 2024). The process begins with construct definition and task development (Wallach et al., 2024), followed by evaluation using multiple human raters and assessment of interrater reliability (Belur et al., 2021).

To enable evaluation at scale, researchers increasingly employ LLM judges that complement human evaluation. While requiring careful validation against human judgments (Li et al., 2024), LLM judges have demonstrated accuracy in educational contexts including standards alignment (Lucy et al., 2024), response scoring (Frohn et al., 2025; Morris et al., 2024), and content refinement (Clark et al., 2025). This dual approach – combining human ground truth with validated LLM evaluation – enables rapid testing and experimentation during development of assessment safeguards.

1.4 Research questions

Intermediate-stakes CBA faces a critical tension: leveraging LLMs' engagement benefits while preventing their tendency to provide inappropriate assistance. This paper addresses this challenge through the following research questions:

1. How accurately can an LLM judge detect inappropriate hints when validated against human expert judgments?

2. Can self-critique mechanisms effectively reduce inappropriate hints in LLM-based CBA?

3. Does self-critique performance generalize from synthetic development data to real-world student conversations?

To address these questions, we develop and evaluate a self-critique mechanism where the LLM evaluates and revises its own responses before delivery. Through two studies – one using synthetic conversations for development and validation, and another using real student pilot data – we demonstrate that prompt engineering can successfully constrain LLM behavior while maintaining the engagement benefits that make CBA valuable for intermediate-stakes assessment.

2 Research

To evaluate whether self-critique can effectively prevent inappropriate hints in CBA interactions, we conducted two complementary studies. Study 1 used synthetic conversations between LLM-simulated students and the assessment chatbot (hereafter "ProctorBot") to develop and validate our self-critique mechanism under controlled conditions. Study 2 validated these findings using real student conversations from high-school math assessment pilots, demonstrating the practical effectiveness of self-critique in authentic assessment contexts.

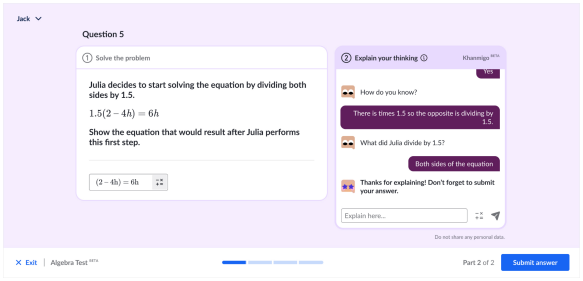

Figure 1: Screenshot of the Explain Your Thinking CBA item type. The student first answers a math problem (left), and then has a conversation about the problem (right) which is designed to assess their conceptual understanding.

2.1 Study 1: Pre-pilot development and validation using synthetic data

Study 1 developed and evaluated the self-critique mechanism under controlled conditions. Using synthetic conversations between LLM-simulated students and ProctorBot, we: (1) collected human expert labels to establish ground truth, (2) validated an LLM judge for detecting inappropriate hints, and (3) conducted an A/B test comparing baseline ProctorBot against a self-critique version.

2.1.1 Methods

Definition of inappropriate hint. For this study, we define an "inappropriate hint" as a ProctorBot response that reveals a concept from the assessment criteria that students are expected to demonstrate. Unlike a response that would draw out a student's thinking and reveal what they know (e.g., a Socratic question), an inappropriate hint would state or strongly hint at one of the criteria concepts – making it difficult to assess what they know. For example, say we wanted to assess if a student understood the concept of inverse operations: If the student solved the problem `1.5x = 3` by dividing, and then ProctorBot asked "How does dividing undo the multiplication?", this would be an inappropriate hint because it reveals the inverse operations concept.

Synthetic data generation. We generated synthetic conversation data using two LLM agents: (1) ProctorBot, designed to assess and question students about their conceptual understanding of math problems, and (2) a student simulator ("StudentBot") designed to express adversarial behaviors (asking for help, expressing uncertainty, refusing to answer) expected to increase the likelihood of inappropriate hints.

Using a Python script to orchestrate conversations between the two agents, we generated 200 synthetic conversations (50 conversations × 4 math problems). Of these, 62 ended early when ProctorBot determined that StudentBot had immediately satisfied the assessment criteria. From the remaining 138 conversations, we systematically extracted 597 test cases at various conversation depths for later experimental use.

The synthetic conversations reflected realistic assessment interactions: StudentBot responses had a median length of 11 words, ProctorBot responses averaged 18 words, and full conversations had a median of 7 turns. To increase response diversity, we varied several StudentBot parameters across simulation runs (see Appendix A).

From this corpus, we sampled 120 ProctorBot responses for human labeling, with some conversations contributing multiple responses from different points in the interaction.

Data labeling and ground truthing. Three subject-matter experts labeled each ProctorBot response as containing an inappropriate hint or not. We presented each response with full context: conversation history, the math problem, assessment criteria, and the inappropriate hint definition.

Initial inter-rater agreement was slight (Fleiss' kappa, denoted κ_F = 0.191 [0.070, 0.314]), with only 53 of 120 items (44%) achieving unanimous agreement. The 67 items with disagreements underwent group discussion and arbitration, resolving 59 cases. This process increased agreement to almost perfect (κ_F = 0.884 [0.798, 0.954]), establishing a reliable ground truth dataset for subsequent analyses.

LLM judge development and validation. To develop an LLM judge capable of detecting inappropriate hints, we tested three prompt variations that differed only in how the target behavior was specified:

1. **Baseline-prompt**: Provided only a simple definition stating that an inappropriate hint "gives away KEY information from the Criteria that has not already been mentioned"
2. **Enhanced-specificity**: Added clarification that "simply mentioning KEY concepts from the Criteria… IS ENOUGH to be considered leading"
3. **Example-based**: Supplemented the baseline definition with six annotated examples (three inappropriate hints, three appropriate responses)

All configurations used GPT-4o with temperature=0 and included the variables in Table 1 as context for the LLM judge. We ran each configuration 20 times on our 120-item dataset to ensure stable estimates, then calculated Cohen's kappa (denoted κ_C) for two-rater agreement and confidence intervals using bootstrap resampling (N=1000) to account for clustering.

Context Element	Description
Problem	The math problem that the student is having a conversation about
StudentAnswer	The student's answer to the problem
Criteria	The assessment criteria
BehaviorDefinition	The definition of inappropriate hints
ConversationHistory	The conversation between student ProctorBot so far
ProctorBotResponse	The ProctorBot response that is being judged, which immediately follows ConversationHistory

Table 1: Context elements provided to the LLM judge.

The enhanced-specificity configuration obtained substantial agreement with ground truth (Landis and Koch, 1977), and had the best balance of performance and simplicity (κ_C = 0.629 [0.611, 0.648]) – outperforming the baseline-prompt (κ_C = 0.553 [0.533, 0.569]), and performing comparably to the significantly more complex example-based prompt (κ_C = 0.612 [0.597, 0.628]). Thus, we decided to use the enhanced-specificity prompt for our implementation of self-critique.

Experiment to evaluate self-critique effectiveness. Having established a reliable automated method for detecting inappropriate hints through our validated LLM judge, we could now evaluate our proposed self-critique intervention at scale. The following experiment tests whether incorporating self-critique into ProctorBot's response generation process can effectively reduce the frequency

of inappropriate hints compared to the baseline system.

From our synthetic dataset of 597 test cases, we identified those with high propensity for inappropriate hints by screening each case 10 times with baseline ProctorBot. This yielded 179 conversation states that produced at least one inappropriate hint (as determined by our LLM judge).

For each of these 179 test cases, we generated responses using both baseline ProctorBot and a self-critique version, then evaluated each response using the LLM judge developed above. The self-critique mechanism employs a two-step process: (1) ProctorBot generates an initial response, then (2) a critic evaluates whether this response inappropriately reveals assessment criteria. If the critic detects an inappropriate hint, it generates a replacement response that avoids revealing assessment criteria. During development, we conducted informal qualitative review of the critic's replacement responses to ensure they maintained pedagogical appropriateness.

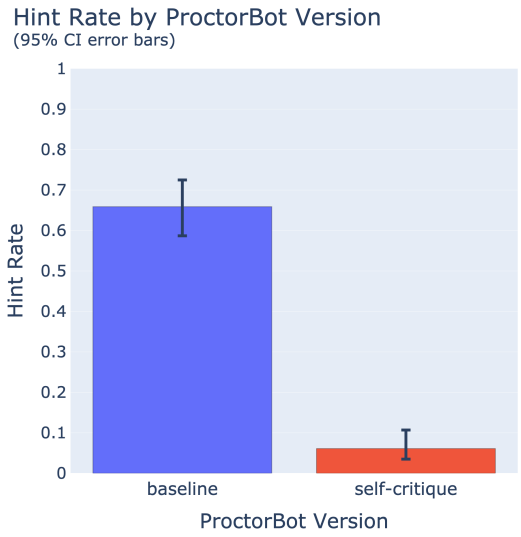

Figure 2: Results of the experiment showing the proportion of inappropriate hints with baseline and self-critique versions of ProctorBot. Self-critique dramatically reduced the rate of inappropriate hints from 65.9% to 6.1% – a 90.7% reduction.

2.1.2 Results

Self-critique dramatically reduced inappropriate hints from 65.9% (118/179) in the baseline to 6.1% (11/179), representing a 90.7% reduction. Figure 2 illustrates this substantial improvement.

To account for the paired nature of our data (same conversation states tested with both versions), we used McNemar's test, which revealed a highly significant difference ($\chi^2 = 101.23$, $p < 0.001$). Of the 111 test cases that showed different outcomes between versions (62% of all cases), 98.2% improved with self-critique: 109 cases changed from producing inappropriate hints to appropriate responses, while only 2 cases showed the opposite pattern.

These findings provided strong evidence for self-critique effectiveness in controlled settings, leading us to validate the approach with real-world data in Study 2.

2.2 Study 2: Post-pilot validation using real-world assessment pilot data

While Study 1 demonstrated self-critique effectiveness with synthetic data, validating this approach with authentic student interactions remained essential.

Study 2 validated the self-critique mechanism in authentic assessment contexts. Using real student conversations from high-school math assessment pilots, we: (1) collected human expert labels to establish ground truth, (2) validated an LLM judge for detecting inappropriate hints, and (3) conducted an A/B test comparing baseline ProctorBot against a self-critique version.

2.2.1 Methods

Data source and sampling. We analyzed conversation data from two high school math assessment pilots (algebra and geometry) conducted between April 14 and May 31, 2025, involving approximately 7,000 students and 9,000 conversations. From this corpus, we sampled 400 conversation states (specific points in conversations where ProctorBot responded), selecting 50 samples from each of eight Common Core standards problems.

To ensure sufficient positive examples given the expected low base rate of inappropriate hints, we performed stratified sampling: we pre-classified 25 examples per problem as likely containing inappropriate hints and 25 as likely not, using GPT-4.1 with the judge prompt from Study 1. We chose GPT-4.1 over GPT-4o for preliminary screening based on its superior agreement with our synthetic ground truth data.

Data labeling and ground truthing. Following the same protocol as Study 1, three subject-matter experts labeled each ProctorBot response. To reduce labeling burden, we employed a tie-break pro-

cess: two raters initially labeled each response, with a third rater resolving disagreements.

Inter-rater agreement (κ_F) was moderate during training (κ_F = 0.428 [0.305, 0.506]) and initially moderate for the main labeling session (κ_F = 0.571 [0.447, 0.682]).[1] Exercise-level analysis revealed that one problem achieved only chance-level agreement (κ_F = -0.004 [-0.389, 0.341]), likely due to ambiguous assessment criteria. Excluding this problem increased overall agreement to substantial (κ_F = 0.669 [0.537, 0.785]).

The final ground truth dataset comprised 350 items, with 82 responses (23.4%) labeled as containing inappropriate hints. Note that this rate reflects our stratified sampling strategy, not the population prevalence in actual student conversations.[2]

LLM judge validation. We validated an LLM judge against the ground truth, testing three models (GPT-4.1, GPT-4o, and GPT-5-mini) and two prompt configurations (baseline and chain-of-thought reasoning). GPT-5-mini without chain-of-thought achieved the strongest agreement with human judgments (κ_C = 0.596 [0.497, 0.688]), reaching a moderate level of agreement (see Appendix B for complete model comparison results).

Confirmatory experiment. To validate whether the self-critique effectiveness observed in Study 1 would generalize to real student conversations, we tested three models (GPT-5-mini, GPT-4.1, GPT-4o) implementing self-critique on our 350 conversation states. We compared these to the original ProctorBot responses from the assessment pilots (baseline), with all responses evaluated using the GPT-5-mini judge.

The self-critique implementation followed the same two-step process as Study 1, with all models operating at temperature=0 (except GPT-5-mini at fixed temperature=1).

2.2.2 Results

Self-critique proved effective with real-world data. All three models showed substantial reductions in inappropriate hints compared to the baseline rate of 27.4% (96/350): GPT-5-mini achieved a 75.0% reduction (to 6.9%), GPT-4.1 a 65.6% reduction (to 9.4%), and GPT-4o a 24.0% reduction (to 20.9%). Figure 3 displays these improvements across models.

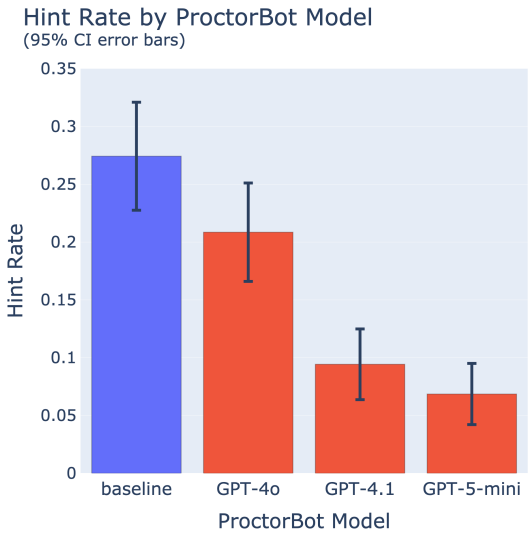

Figure 3: Inappropriate hint rates across model configurations on real-world pilot data. Self-critique implementations achieved reductions ranging from 24% (GPT-4o) to 75% (GPT-5-mini) compared to the null baseline, with all improvements being statistically significant. Error bars represent 95% confidence intervals.

McNemar's test confirmed highly significant differences for all models (all $p < 0.001$, significant after multiple comparison correction). The improvement pattern mirrored Study 1: of discordant pairs, the vast majority (89.1% for GPT-5-mini, 88.9% for GPT-4.1, 79.5% for GPT-4o) changed from inappropriate hints to appropriate responses with self-critique.

3 Conclusions

Our two-study investigation demonstrates that self-critique substantially reduces inappropriate hints in both synthetic and real-world CBA contexts.

Self-critique offers educational institutions a practical, immediately deployable solution for constraining LLM behavior in Conversation-Based Assessment. Organizations can implement this safeguard through simple prompt modifications, avoiding the costs and complexity of model fine-tuning. The technique's accessibility makes it particularly valuable for institutions with limited technical resources.

Our systematic evaluation methodology provides a template for assessing LLM behaviors in educational contexts. We progressed from synthetic to

[1] We excluded tie-break labels from agreement calculations as they are conditionally sampled only when initial raters disagree, violating assumptions for valid kappa statistics.

[2] The true population rate is likely substantially lower, as we deliberately oversampled conversations initially classified as containing inappropriate hints to ensure sufficient positive examples for analysis.

real-world data with rigorous human validation. The significant reductions in inappropriate hints across both studies validate self-critique as an effective safeguard. The same process could be used to attempt to reduce answer giving in other tutor scenarios where providing the answer is not desired. However, important limitations remain. Our evaluation focused specifically on mathematics assessment and hints that reveal assessment criteria. Generalization to other domains and types of assistance requires further investigation. Additionally, we did not systematically evaluate whether self-critique impacts overall response quality or educational value. Our focus remained exclusively on inappropriate hint reduction. While informal qualitative review during development suggested that responses remained pedagogically appropriate, quantifying any trade-offs between constraint effectiveness and response quality remains an open question.

Together with other emerging approaches for quality assurance in educational AI, self-critique offers a targeted solution for constraining LLM outputs through prompt engineering. Our contribution shows that even simple, immediately deployable techniques can meaningfully reduce inappropriate LLM behaviors and advance assessment validity when grounded in rigorous evaluation. As educational institutions navigate the integration of generative AI, this combination of theoretical frameworks, empirical validation, and practical tools will prove essential for maintaining the standards that make assessment meaningful.

4 Appendix A: StudentBot parameter variations

To increase the diversity of synthetic student responses in Study 1, we varied the following StudentBot parameters across simulation runs:

- **Model selection**: GPT-4o and Llama-3.1
- **Initial answer correctness**: Whether StudentBot provided a correct or incorrect answer to the initial math problem
- **Student persona traits**: Anxiety level, communication style (formal vs. informal), patience, and engagement level

These variations ensured that our synthetic dataset captured a range of student behaviors and interaction patterns, improving the robustness of our inappropriate hint detection and self-critique evaluation.

5 Appendix B: LLM judge model comparison results

Complete results from Study 2 LLM judge validation (κ_C with 95% confidence intervals):

- **GPT-5-mini**: κ_C = 0.596 [0.497, 0.688] (without chain-of-thought); 0.551 [0.445, 0.652] (with chain-of-thought)
- **GPT-4.1**: κ_C = 0.422 [0.304, 0.527] (without chain-of-thought); 0.437 [0.303, 0.550] (with chain-of-thought)
- **GPT-4o**: κ_C = 0.195 [0.086, 0.303] (without chain-of-thought); 0.320 [0.200, 0.431] (with chain-of-thought)

References

Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. 2024. A Statistical Framework for Ranking LLM-Based Chatbots. *arXiv preprint*. ArXiv:2412.18407 [stat].

Isaac I. Bejar. 2017. A Historical Survey of Research Regarding Constructed-Response Formats. In Randy E. Bennett and Matthias von Davier, editors, *Advancing Human Assessment: The Methodological, Psychological and Policy Contributions of ETS*, pages 565–633. Springer International Publishing, Cham.

Jyoti Belur, Lisa Tompson, Amy Thornton, and Miranda Simon. 2021. Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. *Sociological Methods & Research*, 50(2):837–865. Publisher: SAGE Publications Inc.

Hannah-Beth Clark, Margaux Dowland, Laura Benton, Reka Budai, Ibrahim Kaan Keskin, Emma Searle, Matthew Gregory, Mark Hodierne, and John Roberts. 2025. Auto-Evaluation: A Critical Measure in Driving Improvements in Quality and Safety of AI-Generated Lesson Resources. *The AI + Open Education Initiative*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv preprint*. ArXiv:2309.11495 [cs].

Hanna Eklöf. 2010. Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, 17(4):345–356.

Bridgid Finn. 2015. Measuring Motivation in Low-Stakes Assessments. *ETS Research Report Series*, 2015(2):1–17.

Scott Frohn, Tyler Burleigh, and Jing Chen. 2025. Automated Scoring of Short Answer Questions with Large Language Models: Impacts of Model, Item, and Rubric Design. In *Artificial Intelligence in Education*, volume VI of *Lecture Notes in Artificial Intelligence*, pages 44–51, Palermo, Italy. Springer.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. Can Large Language Models Detect Errors in Long Chain-of-Thought Reasoning? *arXiv preprint*. ArXiv:2502.19361 [cs].

Cameron R. Jones and Benjamin K. Bergen. 2024. People cannot distinguish GPT-4 from a human in a Turing test. *arXiv preprint*. ArXiv:2405.08007 [cs].

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *arXiv preprint*. ArXiv:2412.05579 [cs].

Li Lucy, Tal August, Rose E. Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. 2024. Mathfish: Evaluating Language Model Math Reasoning via Grounding in Educational Curricula. *arXiv preprint*. ArXiv:2408.04226 [cs].

Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott Crossley. 2024. Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models. *International Journal of Artificial Intelligence in Education*.

Marianne Perie, Scott Marion, and Brian Gong. 2009. Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. *Educational Measurement: Issues and Practice*, 28(3):5–13.

Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. Towards the Pedagogical Steering of Large Language Models for Tutoring: A Case Study with Modeling Productive Failure. *arXiv preprint*. ArXiv:2410.03781 [cs].

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv preprint*. ArXiv:2402.07927 [cs].

Resmi Vijayan and Sunish Vengathattil. 2025. Using the Right Tool: Prompt Engineering vs. Model Tuning. *International Journal of Innovative Science and Research Technology*, pages 274–284.

Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2024. Evaluating Generative AI Systems is a Social Science Measurement Challenge. *arXiv preprint*. ArXiv:2411.10939 [cs].

Seyma N. Yildirim-Erbasli and Okan Bulut. 2023. Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment. *Computers and Education: Artificial Intelligence*, 4:100135.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. *arXiv preprint*. ArXiv:1909.08593 [cs].