Prompting for Teachability: Designing Novice Personas in LLMs for Learning by Teaching Contexts

Sydney Miller
University of Illinois Urbana Champaign
USA
srm16@illinois.edu

Nigel Bosch
Information Science
University of Illinois Urbana Champaign
Urbana, Illinois, USA
pnb@illinois.edu

Abstract

Learning by teaching (LbT) is a well-established instructional framework in which students deepen understanding by explaining material to a peer or tutee. Large Language Models (LLMs) create new opportunities to scale LbT by simulating novice learners, but their default tendency toward expert-like responses risks undermining the tutor's role. This study investigates which prompting strategies most effectively elicit novice-behavior from LLMs in writing-related domains. We generated 30,720 combined prompts across five domains and evaluated three models (Qwen3-235B, Llama 4, Kimi-K2) using both multiple-choice quizzes and short persuasive essays. Outputs were scored on quiz accuracy, essay quality, and essay persuasiveness using an AI-judge rubric. Regression analysis revealed a clear pattern: constraint prompts that explicitly forced error production consistently outperformed persona-, misconception-, and uncertainty-based prompts. Across both quiz and essay outcomes, direct commands to "answer incorrectly" or "get 2–3 wrong" yielded the strongest novice-like behavior, while indirect framings like "don't aim for a perfect score" or "you may guess" diluted the effect. These findings highlight constraint-based prompting as the most reliable strategy, and we argue that constraint directives provide an actionable design pathway for practitioners seeking to integrate LLMs into effective LbT contexts.

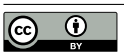
CCS Concepts

• **Applied computing** → Education; E-learning; • **Computing methodologies** → Modeling and simulation; Simulation types and techniques; • **Human-centered computing** → Human computer interaction (HCI); Interaction paradigms; Natural language interfaces.

Keywords

Learning by teaching, Large language models, Simulated students

ACM Reference Format:
Sydney Miller and Nigel Bosch. 2026. Prompting for Teachability: Designing Novice Personas in LLMs for Learning by Teaching Contexts. In *LAK26: 16th International Learning Analytics and Knowledge Conference (LAK 2026), April 27–May 01, 2026, Bergen, Norway*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3785022.3785067



LAK 2026, Bergen, Norway

ACM ISBN 979-8-4007-2066-6/2026/04
https://doi.org/10.1145/3785022.3785067

1 Introduction

Students learn a lot from teaching their peers. Decades of work on learning by teaching (LbT) and peer tutoring show that explaining concepts, anticipating misunderstandings, and responding to questions consolidate understanding and support transfer [5, 8, 9, 16, 22]. These benefits are most evident in collaborative learning contexts [4], and in some well-structured domains that support teachable agents like *Betty's Brain* [25]. We propose that large language models (LLMs) may provide a new opportunity to revisit the LbT paradigm in less-structured domains than what was possible before.

The chat-interface nature of LLMs presents an opportunity for LbT in new contexts since LLMs can flexibly adopt roles via prompting, including the role of a "tutee" that asks questions or makes mistakes [23]. Yet, LLMs also carry their own risks. Unlike hand-engineered teachable agents, LLMs are trained on massive corpora that skew toward fluent, expert-level prose, predisposing them to produce high-competence, authoritative responses by default [14, 19]. In LbT terms, the "expert-like" nature of LLMs risks reproducing the same dominance problem seen in imbalanced peer tutoring [4]: the model does the heavy cognitive lifting, leaving the student little to explain and, therefore, learn from. Without explicit tailoring, the LLM will tend to act like the most competent peer, which would lend to undermining the LbT process rather than enabling it.

However, it may be possible to align LLM behavior with LbT goals via prompting. Prior work shows that detailed, intentional prompts can steer models toward more context-appropriate and role-consistent behavior [14, 18, 23, 28] and that prompting is not a black-box trick but a teachable literacy: structured frameworks and iterative refinement improve immediate output outcomes as well as longer-term metacognitive benefits [26]. Yet most students do not consistently engage in intentional prompting strategies when interacting with LLMs [24], suggesting a practice gap that design can help close by exploring what prompts can be specified in advance to generate LbT experiences from the very first chat turn. Our approach treats prompts as modular by separating identity, tone, behavior, and rules, to counteract the LLM's default expert bias and elicit novice-like responses that keep the human student in the tutor role.

We situate this design within broader conversations about responsible AI in higher education. Scholarship documents risks of bias reproduction, misinformation, and de-skilling, particularly when generative AI is used uncritically [3, 7, 13]. A sociomaterial perspective holds that technologies and practices co-constitute educational activity; thus, the question is not whether to exclude LLMs from learning contexts, but how to shape their participation

strategically through informed constraints and literacies [11, 29]. Designing prompts that *purposefully reduce* model authority in LbT is one such constraint.

Writing education is a particularly strong testbed for this inquiry. Composition theory, broadly understood as the study of writing as both a communicative practice and a cognitive process [20], has long argued that writing is itself a mode of learning since it externalizes thought and reorganizes knowledge for an audience [6, 10]. Both writing and teaching demand articulation, anticipation of audience, and repair of misunderstanding; combining them can create a productive double loop in which teaching reinforces writing and writing reinforces teaching. Writing, therefore, offers an authentic, less-structured context to examine whether prompt engineering can reliably induce novice-like LLM behavior that may support human tutors in a LbT context.

Building on these premises, we investigate:

- To what extent can publicly accessible LLMs that require no fine-tuning be prompted to simulate novice-like behavior in ways that support learning by teaching contexts?
- Which prompting strategies are most effective at eliciting novice-like responses from LLMs?

By addressing these questions, we contribute conceptual and empirical insights into how LLMs might be leveraged as simulated learners, opening new pathways for equitable and accessible LbT opportunities in higher education.

2 Related Work

2.1 Learning by Teaching & LLMs

Learning by teaching (LbT) describes a process in which students deepen their understanding by explaining material to others [8, 9, 22]. LbT is typically practiced in peer tutoring contexts where one student adopts the role of tutor while another acts as the learner or tutee [4]. The LbT paradigm has been shown to increase metacognition, retention, and transfer across disciplines [5, 16].

The learning benefits for tutors arise through both content exposure and from the cognitive work involved in anticipating misconceptions, responding to questions, and articulating the content [8]. However, social and affective dynamics can limit these benefits. Students who perceive themselves as less competent are dominated by peers who take control of the interaction, leading to inequity in the learning process [4]. In other words, the configuration of LbT matters: equity of participation and the tutee's responsiveness both shape how effectively the tutor learns.

Prior systems have explored the use of technology to extend LbT opportunities. For example, platforms such as Betty's Brain simulate teachable agents in computer-based learning environments, enabling students to take on the role of teacher with consistent access to a simulated responsive learner [25]. Studies like Betty's Brain demonstrate that computer-mediated tutees can replicate many of the benefits of peer tutoring while avoiding some of the logistical constraints of pairing students. However, most work has focused on younger learners or narrowly structured domains, leaving open questions about how LbT might be supported for adults in less structured contexts.

As mentioned in the introduction, prompting is crucial to support LbT. LLMs must be carefully guided to behave as novice learners might: fallible, hesitant, and in need of scaffolding. By constraining identity, tone, and behavior through modular prompts, this study aims to counteract the model's expert bias, creating conditions where a human student may retain the role of tutor in a LbT context. This study explores which prompting strategies are most effective for eliciting such novice-like responses, extending the tradition of teachable agents into less-structured domains while mitigating the risks of LLMs dominating the interaction.

2.2 Prompt Engineering & Strategies

As Cain [1] notes, "mastering the process of engineering effective prompts is crucial in fully utilizing the potential of these [GenAI] tools" (p. 49). While the phrase *prompt engineering* may sound technical, Reynolds and McDonell (2021) argue that it is less like writing code and more like writing prose. Since LLMs are trained on massive quantities of human language, prompts operate through tone, framing, and implication [21]. This rhetorical perspective underpins our approach to modular prompt construction, in which identity, tone, behavior, and rules are layered together to guide the LLM toward novice-like behaviors.

As it stands, students rarely engage in the kinds of intentional prompting that yield stronger outputs and deeper learning. Sawalha et al. [24] found that only 40% of students ($n = 54$) regularly crafted multi-part prompts, which was the strategy their study identified as most effective. In other words, most students miss a dual opportunity to both improve the immediate quality of model responses and to cultivate the longer-term learning benefits that come from more intentional prompting. This gap indicates a need for structured, reusable prompt strategies that can lower the barrier to novice-like LLM behavior.

One way to lower this barrier is through prompt patterns, which provide reusable structures that solve recurring interaction challenges [28]. Of particular relevance is the *persona* pattern, which instructs the model to adopt a specific role or identity (e.g., teacher, expert, novice student). The mention of persona-based prompts has become especially common in research speculating the possibilities of AI-augmented teaching and learning, where they are often assumed sufficient for eliciting role-specific responses [2, 15, 17]. Yet this strategy has important limitations. While persona prompts provide a starting point for role-based interactions, they often conflate identity and behavior into a single instruction, limiting control over specific novice-like qualities. A simple prompt such as "act as a novice student" may produce less polished language, but fail to reliably simulate uncertainty, fallibility, or the need for scaffolding. Our approach therefore extends persona prompting by modularizing identity and behavior into distinct components, allowing us to more precisely guide LLMs toward novice-like responses that preserve the tutor's role in LbT interactions.

3 Method

The objective of this study was to investigate which prompting strategies most effectively induce LLMs to adopt the persona of a college-level novice learner, thereby enabling new opportunities for computer-based LbT. Novice-like behavior was defined as response

patterns consistent with learners in the early stages of understanding, such as partial knowledge, inconsistent application of concepts, occasional errors, or hedging language that signals uncertainty.

3.1 Student Behaviors for the LLMs to Simulate

We evaluated novice-like behavior using two complementary measures: (1) multiple choice quiz accuracy, which provided a direct measure of correctness, and (2) essay-based scoring, which captured broader qualities including persuasiveness, writing quality, and degree of "student-likeness" (detailed rubrics in section 3.4). For each prompt, the model was instructed to first answer six multiple-choice questions and then write a short essay. We provided a consistent formatting directive as such within each prompt to enforce a standardized output format:

> "Answer the following multiple choice quiz questions based on the persona instructions in the prompt. Number your answers and provide the letter of your answer, one per line. For example, '1. X.' After the quiz, write a persuasive essay (~250 words) in response to the essay task. Begin the essay on a new line labeled 'Essay:'."

This directive ensured that quiz responses could be automatically scored and essays consistently parsed for evaluation.

To ground these measures in authentic student performance, we drew on a parallel human-subjects pilot study, which had college-level novice writers ($N = 46$) complete a multiple-choice quiz on foundational rhetorical concepts (e.g., identifying ethos, pathos, or logos in a sentence). Rhetorical concepts were used for this pilot since they are both foundational in writing instruction and concrete enough to test novice understanding through multiple-choice items. Preliminary results showed an average score of 4.81/6, which we adopted as a reference point for human novice-level performance. This calibration enabled us to interpret LLM quiz scores relative to real student distributions.

For other content areas in this study (i.e., thesis statements, evidence and support, organization, grammar), no human baseline data were available. In these cases, quiz scores were interpreted comparatively within topic rather than against an absolute threshold. For example, responses scoring around 4/6 were treated as relatively more novice-like than higher-scoring responses, while recognizing that the absence of human baselines limited direct calibration.

With this definition and calibration in place, we then turned to iterative prompt design to test how effectively different strategies elicited novice-like responses.

3.2 Prompt Design

The study progressed through three iterative phases to refine how effectively LLMs could be guided to produce novice-like responses. These phases were exploratory, intended to identify which prompt structures showed the most promise before scaling up to the full experiment.

3.2.1 Round 1: Building Block Prompts. "Building block" prompts tested whether simple modular combinations of prompt components could induce novice-like behavior. We constructed prompts by combining four building blocks:

1. Opening verbs (e.g., "pretend," "act as if," etc.),
2. Identities (e.g., "you are a beginner"),
3. Behavioral traits (e.g., "act confused"), and
4. Rules (e.g., "explain your thinking").

This approach generated a wide variety of prompt formulations, but the responses consistently scored almost perfectly on quizzes. High quiz accuracy indicated that the models were not simulating the kinds of partial understanding of typical novices, which limited their usefulness for our goals. These results motivated us to design prompts that place more deliberate constraints on model behavior in the next phase.

3.2.2 Round 2: Category Prompts. Building on round 1, we developed additional categories of prompt elements to explicitly align with novice-like patterns observed in student data. Specifically, we created five types of prompts:

- Baseline: the general prompts from round 1 (12 sentences per topic)
- Simulating uncertainty: hedging, self-doubt (8 sentences per topic)
- Misconception seeding: flawed reasoning (8 sentences, general across topics)
- Constraint: explicit instruction to make errors at a certain rate (8 sentences, general across topics)
- Combined prompts: one sentence drawn from each of the above categories

From these categories, we sampled 150 unique prompts each, yielding ~750 prompts total. These designs were intended to test whether more focused behavioral instructions, like uncertainty, misconceptions, and constraints, would reduce quiz accuracy and produce more novice-like performance.

3.2.3 Round 3: Deep Dive into Combined Prompts. We further analyzed the combined prompts at the sentence level to understand the effects of each prompt element. Each combined prompt consisted of four sentences: one from each of the four categories (baseline, uncertainty, misconception, constraint). We disaggregated them at the sentence level to isolate the contribution of each element, using a linear model to test how prompt category influenced novice-like behavior.

The outcome of this phase was twofold: (1) it provided evidence about which prompt elements most consistently shaped novice-like responses, and (2) it produced a structured set of combined prompts that could be scaled up in the full study. Together, these iterations supplied both methodological insight and practical materials for the large-scale experiment described in section 3.3.

3.3 Model Execution and Data Collection

Building on the outcomes of round 3, the final study scaled up both in scope and complexity. The experiment was conducted across five writing-related domains: (1) rhetorical strategies, (2) thesis statements, (3) evidence and support, (4) organization and coherence, and (5) grammar. For each topic domain, we generated

6,144 combined prompts (constructed from the four-category design outlined in 3.2, round 2, with $12 \times 8 \times 8 \times 8$ sentences across the categories), paired with a six-question multiple-choice quiz and an essay task. The essay prompt was consistent across all domains.

Performance data were gathered by running the 6,144 combined prompts using three different LLMs, which we chose as options that can be run on-premise (given sufficient hardware), thus avoiding issues of privacy in LbT applications where students could reveal potentially sensitive information during interactions with these models:

1. Qwen3-235B-Instruct (July 2025 release),
2. Llama 4 Maverick, and
3. Kimi-K2

Default sampler settings were used, with temperature = 0.7, unless otherwise specified. Data processing and analysis were implemented in Python 3.12 with the OpenAI API library. The following two steps were performed for each of the three models:

In step 1, we generated outputs for 30,720 unique combined prompts across five writing-related domains (rhetorical strategies, thesis statements, evidence and support, organization, and grammar). Each of the three LLMs (Qwen3-235B-Instruct, Llama 4, and Kimi-K2) was run on this full prompt set, yielding a total of 92,160 prompt-response pairs. Each prompt elicited two outputs: (1) a six-question multiple-choice quiz response and (2) a short essay. For consistency, each model was used to both generate and score its own outputs.

In step 2, we evaluated responses using the rubric described in section 3.4. Each judge produced four scores per output: multiple-choice quiz score, essay quality, essay persuasiveness, and student-likeness.

3.4 AI Judge and Rubric Development

To evaluate LLM output, we developed an automated scoring system referred to as the AI judge. Responses were scored against a four-dimensional rubric:

1. Quiz accuracy: number of correct responses out of six.
2. Essay quality: coherence, clarity, and grammar, rated on a 0–9 scale.
3. Essay persuasiveness: rhetorical effectiveness of the argument, rated on a 0–9 scale.
4. Student-likeness: degree to which the response resembled that of a novice learner (e.g., hedging, partial understanding, minor errors), rated on a 0–9 scale.

Rubric criteria were adapted from established holistic scoring frameworks in writing assessment [27]. Holistic rubrics integrate multiple properties of written expression into broader judgements, rather than isolating micro-level skills, and are widely used in large-scale assessment contexts. In our case, the four rubric dimensions we developed reflect higher-order judgements about writing performance and rhetorical effectiveness.

We manually rated a calibration set of 24 essays across the five domains and compared these human-assigned scores to the AI judge's scores to assess the consistency of the AI-based scoring system. Each essay was independently scored by one human rater and by the AI judge using the same four-dimensional rubric, with dimension scores aggregated to a composite 0–100 scale. The mean human score was 63.3, while the mean AI judge score was 72.1. Although the AI judge's scores were slightly higher on average, the two sets of scores were highly correlated (Pearson's $r = 0.943$, $p < .001$), indicating strong rank-order consistency between human and AI evaluations. Since the full dataset included over 30,000 model-generated essays, this calibrated subsample was used to validate alignment between human and AI scoring rather than computing reliability across all responses.

3.5 Statistical Analysis of Scores

All statistical analyses were conducted in R (version 4.5.1), with multiple linear regression serving as the primary analytic method. Descriptive statistics (means, standard deviations, distributions) were generated to contextualize regression outputs and to provide an overview of performance trends across models and domains. Cross-category comparisons further supported the interpretation of regression coefficients, ensuring that the fine-grained results aligned with broader observable patterns.

We assigned each sentence variant a unique ID factor based on keywords to enable cross-topic comparisons. For example, all prompts beginning with "*Assume you are preparing for a unit test on. . .*" were coded as Sentence 1, ID1, while "*You are a college student learning. . .*" was coded as Sentence 1, ID7. Similarly, directive phrasings in Sentence 4 (e.g., "Answer incorrectly on a few questions. . .", ID1) were distinguished from more passive ones (e.g., "Don't aim for a perfect score. . .", ID2). This ID system allowed us to examine the rhetorical and stylistic differences within each sentence slot. We set the reference level to whichever ID was closest to the mean, so that coefficients represented the change in score due to a particular sentence relative to the mean.

4 Results

All three regression models were statistically significant, indicating that sentence choices explained a meaningful proportion of the variance in outcomes. We examined sentence variants that were (a) statistically highly significant ($p < .01$) across all three models and (b) directionally consistent (i.e., all positive or all negative coefficients) across all three models. In terms of quizzes, this filtering yielded 15 robust effects that worked across LLMs.

Table 1 shows the sentence variants that significantly influenced quiz performance. In general, explicit constraint-type sentences had the largest effects.

Table 2 highlights the corresponding results for essay outcomes. Since essay *quality* and *persuasiveness* were conceptually related and often produced overlapping sets of significant sentence variants, we averaged these two dimensions to create a single composite outcome. Only sentence variants that were statistically significant ($p < .01$) and directionally consistent across models for both quality and persuasiveness were included prior to averaging.

5 Discussion

A clear pattern emerged across both quiz and essay outcomes: the degree of directiveness in sentence phrasing consistently shaped effectiveness. The strongest effects overall were observed in Sentence 4 (constraint). Direct commands such as "Answer incorrectly on a few questions" (ID1) and "Try to get 2–3 wrong. . ." (ID7) yielded the

Table 1: Statistically significant sentence variants for quiz outcomes

			Coefficient (influence on quiz score)			
Sentence Type	ID	Sentence Prefix	Qwen3-235B	Llama 4	Kimi-K2	Mean
Baseline	7	"You are a college student learning…"	0.496	0.213	0.256	0.321
Baseline	11	"You are taking an introductory class focused on…"	0.320	0.189	0.228	0.246
Baseline	12	"You are a student…"	0.497	0.142	0.314	0.318
Uncertainty	1	"You assume…"	0.394	0.323	0.482	0.400
Uncertainty	2	"You believe…"	0.200	0.315	0.150	0.222
Uncertainty	4	"You've heard…"	-0.603	-0.153	-0.154	-0.304
Misconception	3	"You are not very confident…"	-0.260	-0.183	-0.185	-0.209
Misconception	5	"You may guess on a few answers."	0.255	0.106	0.370	0.244
Misconception	6	"You might not remember everything you've been taught."	0.319	0.098	0.172	0.197
Misconception	8	"You sometimes mix up the definitions."	-0.322	-0.239	-0.455	-0.339
Constraint	1	"Answer incorrectly on a few…"	-2.537	-1.196	-2.269	**-2.001**
Constraint	2	"Don't aim for a perfect score…"	1.667	0.675	1.458	**1.266**
Constraint	3	"Don't answer everything perfectly…"	1.610	0.664	1.372	**1.215**
Constraint	4	"Include 2–3 errors to show you are still learning."	0.188	0.436	0.655	0.426
Constraint	7	"Try to get 2 or 3 answers wrong."	-1.871	-0.297	-1.300	**-1.156**

[a] Coefficients from regression models for Qwen, Llama, and Kimi. All values included were statistically significant ($p < .01$) across models and directionally consistent. Negative coefficients indicate more effective prompts (lower quiz scores), while positive coefficients indicate less effective prompts (higher quiz scores). Average estimates are reported in the final column, with values of magnitude $\geq \pm 1.0$ bolded for emphasis.

Table 2: Statistically significant sentence variants for essay outcomes

			Average Estimate of Essay Quality & Persuasiveness Scores			
Sentence Type	ID	Sentence Prefix	Qwen3-235B	Llama 4	Kimi-K2	Mean
Baseline	6	"Take this quiz as if you are a novice."	-0.446	-0.371	-0.618	-0.478
Baseline	8	"You are reviewing…"	0.102	0.200	0.823	0.375
Baseline	9	"You are studying…"	0.149	0.229	1.146	**0.508**
Misconception	3	"You are not very confident…"	-0.219	-0.250	-0.493	-0.321
Misconception	5	"You may guess on a few answers."	0.118	0.213	0.210	0.180
Constraint	2	"Don't aim for a perfect score…"	0.318	0.538	0.681	**0.513**
Constraint	3	"Don't answer everything perfectly…"	0.290	0.573	0.917	**0.593**
Constraint	4	"Include 2–3 errors to show you are still learning."	0.148	0.283	0.423	0.285
Constraint	6	"Make a few mistakes that someone new to persuasive writing might make."	-0.360	-0.358	-1.306	**-0.675**
Constraint	7	"Try to get 2 or 3 answers wrong."	0.099	0.566	0.878	**0.514**

[a] Average regression coefficients for sentence variants that were statistically significant ($p < .01$) and directionally consistent across Qwen, Llama, and Kimi on essay quality and persuasiveness estimates. Reported values represent the mean of quality and persuasiveness estimates. Negative coefficients indicate more effective prompts (lower scores), while positive coefficients indicate less effective prompts (higher scores). Bolded values in the Mean column mark averages of magnitude $\geq \pm 0.5$.

most negative coefficients across models, meaning they were the most effective at eliciting fallible, novice-like behavior. By contrast, indirect framings like "Don't aim for a perfect score" (ID2) and "Don't answer everything perfectly" (ID3) produced large positive coefficients, suggesting that hedged or negative instructions were less effective. These findings indicate that directive, prescriptive instructions were most effective at producing learner-like behavior, while indirect instructions often diluted this effect.

This trend extends beyond constraint prompts. In Sentence 3 (misconception seeding) prompts that directly stated weaknesses (e.g., "You are not very confident", ID3; "You sometimes mix up definitions", ID8) consistently reduced performance, while hedged

variants (e.g., "You may guess on a few answers", ID5; "You might not remember everything you've been taught", ID6) produced more positive outcomes. Similarly, in Sentence 2 (simulated uncertainty), internal cognitive framings like "You assume" (ID1) and "You believe" (ID2) yielded positive outcomes, whereas the more externally grounded "You've heard" (ID4) consistently produced negative coefficients. These results indicate that across sentence types, the rhetorical stance of phrasing as either directive or indirective is crucial in shaping outcomes.

Essay outcomes reinforced this pattern. In Table 2, the constraint sentence type again accounted for the largest share of effects: 5 out of the 10 significant and directionally consistent sentence variants came from constraint sentences. Moreover, the sentences with the strongest magnitudes (whether positive or negative) were also most often drawn from the constraint category. When comparing across outcomes (Tables 1 and 2), several sentence variants appeared in both the quiz and essay analyses. These included misconceptions 3 and 5, as well as four total constraint sentences: constraint 2, 3, 4, and 7. The overlap suggests that these constraint framings in particular produce consistent effects across both quiz and essay tasks, making this sentence type especially prominent across the study.

These findings should be interpreted as an exploratory step, as the study examines simulated novice behavior and relies on automated scoring for large-scale comparison rather than direct measurement of human learning outcomes.

5.1 Constraint Prompts as the Most Effective Strategy

The most striking result across both quiz and essay outcomes was the dominance of constraint prompts. While much of the prior work on education-based prompt engineering has emphasized persona-based prompting as an effective strategy, our findings show that such role framing is comparatively weak for LbT purposes. Instead, the most effective way to elicit fallible, learner-like behavior was to hard-code error production through explicit constraints (e.g., "Try to get 2–3 wrong"). Constraint prompts accounted for the largest share of robust effects and consistently produced the strongest magnitudes across both quiz and essay measures. In other words, rather than asking the model to *be* a novice, it is more reliable to explicate how the model should *perform* like a novice through clear forced error directives.

5.2 Pedagogical Implications

For educators and researchers designing LbT environments with LLMs, these findings suggest that persona-based role framing alone is insufficient for eliciting novice-like behavior. Instead, constraint-based prompts that explicitly require imperfect performance (e.g., specifying a small number of incorrect responses) provide a more reliable mechanism for preserving students' tutor roles in LbT activities. Persona elements may still support context and tone, but our results indicate they are most effective when paired with explicit behavioral constraints rather than used in isolation.

6 Conclusion

This study examined whether publicly accessible LLMs can be steered into novice-like roles that support learning by teaching (LbT), and which prompt engineering strategies most effectively licit such behavior. Across both quiz and essay outcomes, we found that directiveness in phrasing—and, in particular, constraint prompts that hard-code fallibility—consistently outperformed other approaches. Thus, we suggest that researchers designing LbT activities should pair role framing, or other existing prompting patterns of favor, with explicit behavior constraints. Future research should examine newer models and modalities, extend analysis to longer and more realistic LbT interactions to test the durability of prompt effects, and, most crucially, measure human learning gains when students teach constraint-guided LLM tutees in a LbT context. As generative AI becomes increasingly embedded in higher education, constraint-based prompting can act as a more reliable approach to shaping LLMs into teachable peers that sustain, rather than undermine, the LbT process.

References

[1] William Cain. 2024. Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education. *TechTrends* 68, 1 (January 2024), 47–57. https://doi.org/10.1007/s11528-023-00896-0

[2] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns* 6, 6 (June 2025), 101260. https://doi.org/10.1016/j.patter.2025.101260

[3] Debby R. E. Cotton, Peter A. Cotton, and J. Reuben Shipway. 2024. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International* 61, 2 (March 2024), 228–239.

[4] Amy Debbané, Ken Jen Lee, Jarvis Tse, and Edith Law. 2023. Learning by Teaching: Key Challenges and Design Implications. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 1–34. https://doi.org/10.1145/3579501

[5] David Duran. 2017. Learning-by-teaching. Evidence and implications as a pedagogical mechanism. *Innovations in Education and Teaching International* 54, 5 (September 2017), 476–484. https://doi.org/10.1080/14703297.2016.1156011

[6] Janet Emig. 1977. Writing as a mode of learning. *College Composition & Communication* (1977).

[7] Emilio Ferrara. 2023. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *FM* (November 2023). https://doi.org/10.5210/fm.v28i11.13346

[8] Logan Fiorella and Richard E. Mayer. 2013. The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology* 38, 4 (October 2013), 281–288. https://doi.org/10.1016/j.cedpsych.2013.06.001

[9] Logan Fiorella and Richard E. Mayer. 2016. Eight Ways to Promote Generative Learning. *Educational Psychology Review* 28, 4 (2016), 717–741.

[10] Linda Flower and John R Hayes. 1981. A Cognitive Process Theory of Writing. *College Composition land Communication* (1981).

[11] Halvdan Haugsbaken and Marianne Hagelia. 2024. A New AI Literacy For The Algorithmic Age: Prompt Engineering Or Eductional Promptization? In *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)*, April 16, 2024. IEEE, Halden, Norway, 1–8. https://doi.org/10.1109/ICAPAI61893.2024.10541229

[12] J. Shieh. 2023. Best practices for prompt engineering with the OpenAI API. *OpenAI Help Center*. Retrieved September 9, 2025 from https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api

[13] Kokil Jaidka, Tsuhan Chen, Simon Chesterman, Wynne Hsu, Min-Yen Kan, Mohan Kankanhalli, Mong Li Lee, Gyula Seres, Terence Sim, Araz Taeihagh, Anthony Tung, Xiaokui Xiao, and Audrey Yue. 2025. Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy. *Digit. Gov.: Res. Pract.* 6, 1 (March 2025), 1–15. https://doi.org/10.1145/3689372

[14] Meiqing Jin, Liam Dugan, and Chris Callison-Burch. 2025. Controlling Difficulty of Generated Text for AI-Assisted Language Learning. https://doi.org/10.48550/arXiv.2506.04072

[15] Nils Knoth, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister. 2024. AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence* 6, (June 2024), 100225. https://doi.org/10.1016/j.caeai.2024.100225

[16] Andreas Lachner, Leonie Jacob, and Vincent Hoogerheide. 2021. Learning by writing explanations: Is explaining to a fictitious student more effective than self-explaining? *Learning and Instruction* 74, (August 2021), 101438. https://doi.org/10.1016/j.learninstruc.2020.101438

[17] Daniel Lee and Edward Palmer. 2025. Prompt engineering in higher education: a systematic review to help inform curricula. *Int J Educ Technol High Educ* 22, 1 (February 2025), 7. https://doi.org/10.1186/s41239-025-00503-7

[18] Leo S. Lo. 2023. The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship* 49, 4 (July 2023), 102720. https://doi.org/10.1016/j.acalib.2023.102720

[19] Ali Malik, Stephen Mayhew, Chris Piech, and Klinton Bicknell. 2024. From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation. https://doi.org/10.48550/arXiv.2406.03030

[20] Stephen M. North. 1987. *The Making of Knowledge in Composition: Portrait of an Emerging Field*. Boynton/Cook Publishers.

[21] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, May 08, 2021. ACM, Yokohama Japan, 1–7. https://doi.org/10.1145/3411763.3451760

[22] Rod D. Roscoe and Michelene T. H. Chi. 2008. Tutor learning: the role of explaining and responding to questions. *Instructional Science* 36, 4 (2008), 321–350.

[23] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-Context Impersonation Reveals Large Language Models' Strengths and Biases.

[24] Ghadeer Sawalha, Imran Taj, and Abdulhadi Shoufan. 2024. Analyzing student prompts and their effect on ChatGPT's performance. *Cogent Education* 11, 1 (December 2024), 2397200. https://doi.org/10.1080/2331186X.2024.2397200

[25] James R. Segedy, John S. Kinnebrew, and Gautam Biswas. 2015. Using Coherence Analysis to Characterize Self-Regulated Learning Behaviours in Open-Ended Learning Environments. *Journal of Learning Analytics* 2, 1 (May 2015), 13–48. https://doi.org/10.18608/jla.2015.21.3

[26] Mo Wang, Minjuan Wang, Xin Xu, Lanqing Yang, Dunbo Cai, and Minghao Yin. 2023. Unleashing ChatGPT's Power: A Case Study on Optimizing Information Retrieval in Flipped Classrooms via Prompt Engineering. *IEEE Trans. Learning Technol.* 17, (October 2023), 629–641. https://doi.org/10.1109/TLT.2023.3324714

[27] S.C. Weigle. 2002. *Assessing Writing*. Cambridge UniversityPress.

[28] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. https://doi.org/10.48550/arXiv.2302.11382

[29] Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int J Educ Technol High Educ* 16, 1 (December 2019), 39. https://doi.org/10.1186/s41239-019-0171-0