CPT-Agent: A Cognitive Process Theory-driven Framework for Student Simulation in Writing Development

Yuhan Chen[1], **Zizhuo Shen**[2], **Miaomiao Cheng**[1], **Xu Han**[1],
Jiefu Gong[2], **Shijin Wang**[2], **Wei Song**[1*]
[1]Information Engineering College, Capital Normal University, Beijing, China
{2241002031, miaomiao, hanxu, wsong}@cnu.edu.cn
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
{zzshen, jfgong, sjwang3}@iflytek.com

Abstract

Simulating student writing behavior offers a promising pathway to scalable feedback evaluation and teacher training. However, existing LLM-based approaches tend to model overly capable learners who readily understand and over-apply feedback, resulting in pedagogically implausible behavior. In this work, we introduce pedagogical realism as a guiding principle for student writing simulation, emphasizing bounded cognition, selective feedback comprehension, and developmentally plausible learning processes. To operationalize this idea, we propose CPT-Agent, a cognitively grounded framework that decouples cognitive ability from writing proficiency and models their interaction during writing and revision. CPT-Agent combines probabilistic modeling of cognitive development, proficiency-controlled text generation, and structured memory for skill accumulation. Experiments show that it (1) produces clearly distinguishable proficiency levels, (2) generates cognitively plausible revisions consistent with instructional theories, and (3) achieves strong agreement with expert judgments in evaluating feedback quality. These results highlight the importance of modeling cognitive constraints in LLM-based student simulation and demonstrate the potential of pedagogically realistic agents for automated feedback assessment and teacher development.

1 Introduction

Large language models (LLMs) show great promise for educational simulation, particularly in addressing the scarcity of teacher-student interaction data and providing risk-free practice environments for novice teachers (Markel et al., 2023; Pan et al., 2025; Bhowmik et al., 2024). In writing education, providing formative feedback on student writing is essential for writing development and is a core competency of effective writing instruction (Graham et al., 2011). Accurately simulating student writing and revision behavior in response to feedback can support both feedback quality evaluation and teacher training.

However, simulating realistic student writing behavior in a controllable and interpretable manner remains an open challenge. While using a single strong model for role-playing (Xu et al., 2024) is convenient, this approach presents several inherent limitations. First, the black-box nature of these models restricts fine-grained control over internal cognitive processes. It is difficult to specify how the model interprets or applies feedback. Second, despite their strong linguistic capabilities, their generic architecture may struggle to authentically represent the developmental nuances and writing characteristics associated with different proficiency levels (Li et al., 2025). As a result, simulated students' responses may not align with empirically grounded patterns of human learning.

To address these challenges, we propose **CPT-Agent**, a novel simulation framework grounded in the Cognitive Process Theory (CPT) of Writing (Flower and Hayes, 1981). The key idea is to *decouple cognitive ability from writing proficiency*, allowing each to be modeled and controlled independently. As shown in Figure 1, CPT-Agent instantiates this idea through three modules:

- **Planner**: Filters feedback according to grade-aware cognitive modeling.
- **Translator**: Generates and revises text at a specified proficiency level.
- **Memory**: Accumulates writing skills from comprehended feedback over time.

Our experiments focus on evaluating the pedagogical realism in three dimensions: Writing proficiency distinctness, cognitive consistency in revision, and developmental authenticity. Both automated and human evaluations show that CPT-Agent outperforms competitive LLM baselines.

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18596–18616

Furthermore, we demonstrate the practical utility of CPT-Agent for evaluating teacher feedback quality. A simulation-based ranking of feedback from novice teachers shows promising agreement with expert judgments. Interviews with these novice teachers confirm the framework's value, indicating that the diagnostic insights generated by CPT-Agent provide actionable guidance for refining their feedback practices.

In summary, our contributions are as follows:

- We identify a key limitation of existing LLM-based student writing simulations that they tend to model overly capable writers who readily understand and over-apply feedback, exhibiting unrealistically strong performance. We introduce *pedagogical realism* as a guiding principle, emphasizing the need to model bounded cognition, selective feedback comprehension, and developmentally plausible learning behaviors.

- We propose CPT-Agent, which operationalizes pedagogical realism by decoupling cognitive ability from writing proficiency. In particular, we model cognitive development as an ordered latent variable and instantiate it with an ordered logit model, enabling smooth and probabilistic transitions between developmental stages. Combined with proficiency-specific fine-tuning and a structured memory module, this design allows controlled and interpretable simulation of student behavior.

- Through both automated and human evaluations, we show that CPT-Agent (1) generates essays that accurately reflect target proficiency levels, (2) produces cognitively plausible revisions consistent with instructional theories, and (3) shows promising alignment with expert human judgment when used to assess novice teachers' feedback.

2 Related Work

2.1 Large Language Models in Education

Recent advances in LLMs have significantly impacted education (Wang et al., 2025a). AI-driven teaching assistants have been developed to alleviate teachers' workload and enhance instructional efficiency (Chu et al., 2025; Pinto et al., 2023; Doughty et al., 2024; Jury et al., 2024). Intelligent tutoring systems have also shown promising results across a range of disciplines (Han et al., 2023; Sonkar et al., 2023; Liu et al., 2024). For example, ChatTutor (Chen et al., 2024) delivers personalized learning experiences, while EduBot (Li et al., 2024) enables the creation of subject-specific teaching assistants tailored to curriculum needs.

In the context of writing education, LLMs have demonstrated significant capabilities as automated essay scorers and feedback generators (Liu et al., 2025; Bui and Barrot, 2025; Han et al., 2023; Stahl et al., 2024). LLM-based applications such as AI Personas (Benharrak et al., 2024) allow writers to receive personalized feedback.

Despite these advances, most existing work focuses on scaffolding writers rather than helping teachers practice and refine their feedback-giving skills (Markel et al., 2023; Jin et al., 2025; Wang et al., 2025a). Our work addresses this gap by introducing a framework for automated evaluation of feedback quality through student simulation.

2.2 Student Simulation

Simulating student behaviors can effectively address the scarcity of teacher-student interaction data (Zhao et al., 2023; Xu and Zhang, 2023), while also providing a pathway for training novice teachers (Markel et al., 2023; Pan et al., 2025; Jin et al., 2025) and iterating educational systems (Li et al., 2024). Most existing approaches employ LLMs initialized with predefined student profiles to emulate learner behaviors (Benedetto et al., 2024; Liu et al., 2024; Yuan et al., 2025; Srivatsa et al., 2025). Some works such as Agent4Edu (Gao et al., 2025b) and GenMentor (Wang et al., 2025b) integrate multiple modules to support more comprehensive educational simulations. However, these approaches often fail to capture students' cognitive states and face issues of blurred capability boundaries in specific role-playing scenarios (Li et al., 2025).

To address these, recent studies have proposed filtering mechanisms (Li et al., 2025), though such methods incur substantial cost. AlgoBo (Jin et al., 2024) employs a reflect-respond pipeline to model learners' knowledge states, but it remains constrained to algorithm learning tasks with objective solutions. These limitations highlight the need for cognitively grounded and controllable simulation frameworks. In this work, we decompose the writing process into three specialized modules inspired by CPT, which enables fine-grained adjustment of learner capabilities while ensuring realism across diverse student profiles.

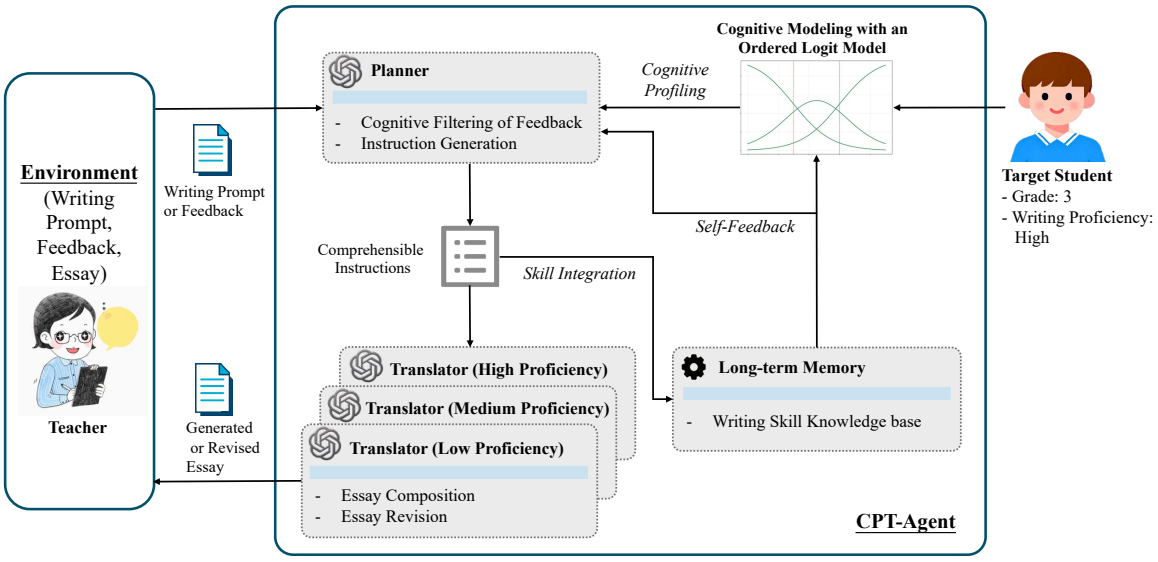


Figure 1: The CPT-Agent architecture comprises three core modules: A Planner, a Translator, and a Memory. The system is initialized with a student's grade and predefined writing proficiency. From this, a cognitive profile is derived using an ordered logit model, based on the grade level and the writing skills in the Memory. The Planner then filters teacher feedback to extract comprehensible writing instructions. These filtered instructions subsequently guide the Translator in composing or revising essays and update the Memory to drive writing skill development.

3 The Student Writing Simulation Task

This work investigates the simulation of student writing processes using LLM-based agents. We focus on essay writing and formulate the agent's behavior with the function

$$y = \texttt{Agent}(X, S),$$

where X is the task environment (a writing prompt or a revision suggestion), S is a configurable student profile and y denotes the resulting essay.

A fundamental design principle is pedagogical realism. The simulated agent is designed not to produce ideal text, but to exhibit characteristic writing behaviors of students at specific grade and proficiency levels. To serve this role effectively, agents must exhibit the following capabilities:

- **Writing capability distinctness**: The ability to generate text that demonstrates distinct and consistent levels of writing proficiency. For example, on the same topic, a low-proficiency agent should produce noticeably simpler vocabulary and weaker organization than a high-proficiency agent.
- **Cognitive consistency in revision**: The property that an agent's responses to feedback should be logically constrained and plausible for its designated grade level. For instance, a second-grade student should struggle with high-level structural feedback, as such skills typically emerge later in development.
- **Developmental authenticity**: The manifestation of a realistic progression in writing skills through iterative learning interactions, mirroring human learning trajectories rather than exhibiting implausible jumps in quality.

4 CPT-Agent

4.1 Overview

Inspired by the Cognitive Process Theory of Writing (Flower and Hayes, 1981), which conceptualizes writing as a goal-directed, recursive activity (See Appendix A), we propose CPT-Agent, whose architecture and workflow are depicted in Figure 1. It has the following modules:

- **Planner**: Acts as the central controller of the writing process, modeling cognitive ability for feedback comprehension, coordinating with memory, and generating adaptive instructions.
- **Translator**: Translates task context and Planner guidance into concrete essay content, reflecting the agent's designated writing proficiency level and adhering to instructions to ensure consistency with assigned skill levels.
- **Memory**: Maintains a dynamic knowledge base of writing skills acquired from feedback,

enabling long-term development.

The core innovation of CPT-Agent lies in its principled decoupling of internal cognitive simulation from external writing proficiency, implementing the theoretical distinction between task representation and process orchestration (Hayes, 2012). This separation is further enriched by the dynamic Memory module that evolves through the Planner's activities. Memory not only archives acquired knowledge but also models cognitive growth, thereby continuously supplementing the Planner's strategic reasoning.

4.2 The Planner

The Planner is formulated as $F_c = \texttt{P}(X, S_c, F)$, where X denotes the task environment; S_c represents the student's cognitive profile, which comprises the student's grade level and cognitive settings for feedback comprehension; F is the provided feedback, which is processed into a set of comprehensible suggestions aligned with S_c, and then combined with writing knowledge to produce specific, concrete guidance F_c. In the case of initial writing, the skill knowledge base K in the memory takes the role of F and serves as self-feedback.

The rationale for this module is that students may not always fully understand complex feedback. A second-grader, for example, can process spelling corrections but not abstract advice about thematic coherence. The Planner simulates this selective comprehension, capturing only the aspects of feedback that the student can genuinely grasp and utilize for subsequent revisions.

4.2.1 Cognitive Modeling

Informed by research on writing development (Graham and Perin, 2007) and cognitive overload (Sweller, 1988), we initially categorize the comprehension of the writing feedback into three distinct levels (*Low, Moderate, Strong*) according to the grade levels. However, fixed levels alone fail to account for the probabilistic nature of student cognitive development. As argued by Siegler (1998), cognitive development is characterized not by a discrete staircase of mastery, but by overlapping waves of strategies where multiple levels of competence coexist and compete.

We move beyond deterministic prompting and implement an *Ordered Logit Model* for the Planner module. We define the model $\texttt{OLM}(\alpha_i, g, \theta_1, \theta_2, \beta)$, which characterizes the stochastic transition between cognitive levels. Let $Y \in \{1, 2, 3\}$ denote the cognitive level, where $Y = 1, 2$ or 3 corresponds to *Low, Moderate*, or *Strong*. The model assumes the existence of threshold parameters θ_1 and θ_2 (where $\theta_1 < \theta_2$), a coefficient β (representing the effect of grade level g on the cognitive level). To account for individual skill variance, we introduce a personalized parameter $\alpha_i \in [0, 1]$, which quantifies the additional cognitive capability derived from the specific writing skills stored in the agent's memory.

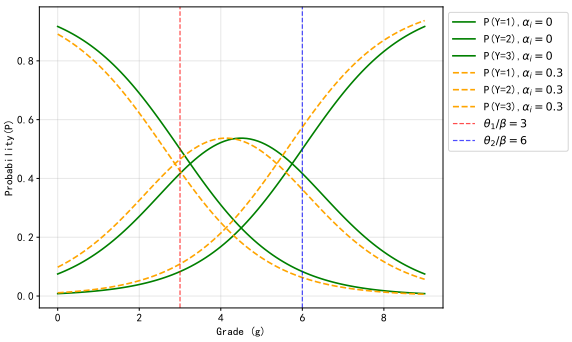


Figure 2: The probability curves of the ordered logit model with $\theta_1 = 2.4, \theta_2 = 4.8, \beta = 0.8$, illustrating how the probability of higher cognitive levels increases with grade level, adjusted by individual capacity α_i.

The cumulative probability is defined as:

$$P(Y \leq j \mid g) = \sigma(\theta_j - \beta g - \alpha_i), \quad for \quad j = 1, 2, \tag{1}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic cumulative distribution function. Here, $\beta > 0$ ensures that higher grade levels correlate with advanced cognitive levels, while θ_1 and θ_2 control the switching points between levels. The probabilities for each level are then:

$$\begin{aligned} P(Y = 1 \mid g) &= \sigma(\theta_1 - \beta g - \alpha_i) \\ P(Y = 2 \mid g) &= \sigma(\theta_2 - \beta g - \alpha_i) \\ &\quad - \sigma(\theta_1 - \beta g - \alpha_i) \\ P(Y = 3 \mid g) &= 1 - \sigma(\theta_2 - \beta g - \alpha_i). \end{aligned} \tag{2}$$

As illustrated in Figure 2, the model exhibits characteristic S-shaped transitions between cognitive levels. The parameters ($\theta_1 = 2.4, \theta_2 = 4.8, \beta = 0.8$) are calibrated so that the crossover from *Low* to *Moderate* occurs around grades 3–4 and from *Moderate* to *Strong* around grades 6–7, consistent with documented developmental milestones in writing cognition (Graham and Perin, 2007; Piaget, 1971).

We estimate the personalized parameter α_i through a prompt-based assessment of the writ-

ing skills in the agent's memory. For example, a grade-5 student with $\alpha_i = 0.3$ yields probabilities $P(Y{=}1) \approx 0.13$, $P(Y{=}2) \approx 0.49$, $P(Y{=}3) \approx 0.38$. This characterizes a predominantly *Moderate* cognitive profile that retains a substantial likelihood of *Strong* comprehension, effectively capturing the stochastic nature of student development.

4.2.2 Instruction Instantiation

For a given g and the estimated α_i for an agent, we derive the level probabilities $\{P(Y = j)\}_{j=1}^{3}$ via the Ordered Logit Model (Equation 2). Building on evidence that LLMs can effectively reason with probabilistic information (Zhang et al., 2024), we combine these probabilities with cognitive level descriptions to instantiate a specific profile S_c, which dictates the instructions used by the Planner to modulate raw feedback F into comprehensible feedback F_c. By performing this *cognitive filtering*, the framework simulates a key psychological reality that students do not passively record feedback. Instead, they interpret and internalize information through the selective lens of their current developmental stage. Detailed prompt templates for this process are provided in Appendix B.

4.3 The Memory

The Memory is defined as $\texttt{M} = (K, \texttt{KI})$. K is a writing skill knowledge base, which organizes basic skills (e.g., vocabulary and grammar) and genre-related skills (e.g., narration and argumentation) into a hierarchical structure. `KI` denotes the Knowledge Integrator, responsible for distilling and integrating skills from the comprehensible feedback F_c into K. We formulate this update process as $K = \texttt{KI}(K, S, F_c)$. By incorporating the student profile S, `KI` facilitates cognitively-aligned knowledge acquisition, aiming to simulate individual variance in the ability to abstract and generalize strategies. This ensures that the evolution of the knowledge base remains consistent with the student's developmental stage, transforming task-specific suggestions into transferable writing strategies. The structure of the writing skill knowledge base and detailed prompts for the Memory are provided in Appendix C.

4.4 The Translator

The Translator `T` emulates essay generation based on a configurable proficiency level and the Planner's instructions. The process is formally defined as $y = \texttt{T}(X, S_w, F_c)$, where X is the environment; S_w represents the student's writing proficiency profile; F_c denotes the Planner's instruction; and y is the output essay. Therefore, the quality of output essays is jointly determined by the Translator's inherent proficiency and the Planner's instructions.

4.4.1 Writing Proficiency Levels

Writing research typically classifies students' writing proficiency into *Low*, *Medium*, and *High* levels based on a combination of writing performance, cognitive skills, and assessment results (Harris and Graham, 1996; Graham and Perin, 2007). Following this classification, we operationalize the writing proficiency profile S_w across two interdependent dimensions: foundational composition proficiency and revision proficiency.

The former governs the agent's baseline linguistic performance, while the latter dictates its capacity to apply feedback during the revision phase. This dual-perspective design is theoretically motivated by observation in writing education, where a student's composition proficiency and their ability to utilize feedback are positively correlated. Higher-proficiency students extract greater utility from feedback, whereas lower-proficiency learners often struggle to deal with suggestions (Hyland and Hyland, 2019).

To capture this interdependence, we utilize a proficiency-level-dependent instruction set. This ensures the agent's response to feedback remains internally consistent with its underlying linguistic capability. Detailed descriptions of each writing proficiency level and their corresponding instruction sets are provided in Appendix D.

4.4.2 Proficiency Modeling via Fine-tuning

While the prompt-based method offers a straightforward implementation, our empirical observations reveal that this approach fails to reliably generate essays with clearly distinguishable proficiency levels. In particular, generic LLMs often struggle to authentically replicate the distinctive linguistic patterns and developmental constraints characteristic of low-proficiency students. To address this, we develop specialized models for the Translator through a targeted, two-stage fine-tuning approach.

In the first stage, we fine-tune a base model on a curated dataset of proficiency-rated essays, establishing the foundational capabilities for generating text at distinguishable writing proficiency levels. We train separate model instances for each writing proficiency level to ensure specialized adaptation.

The second stage employs knowledge distillation from GPT-4o to create a targeted dataset of feedback-revision pairs. For each writing proficiency level, we use the corresponding first-stage essays, GPT-4o then simulates both feedback generation and student feedback processing. To maintain feedback application alignment, GPT-4o was prompted to generate feedback using grade-specific linguistic constraints, ensuring the lexical complexity and syntactic structures were tailored to the target proficiency level. This process yields a dataset in ⟨original essay, feedback, revised essay⟩format. We use this dataset to further fine-tune the proficiency-specific models, enhancing their capacity to simulate revision behaviors. Fine-tuning details are provided in Appendix D.4.

During simulation, the Translator selects the proficiency-specific model based on the target student profile and uses the Planner's instruction F_c as input to generate revised essays.

5 Experiments

We examine the pedagogical realism in student writing simulation through the following research questions:

(1) Can CPT-Agent effectively simulate distinguishable and consistent writing proficiency levels across diverse student profiles?

(2) Do simulated students demonstrate cognitively consistent responses to feedback that align with established educational theories and enable rational learning progression?

(3) Can the simulation framework reliably evaluate feedback quality using simulated students?

5.1 Experimental Settings

CPT-Agent uses Qwen2-7B-Instruct (Qwen2-7B) (Team, 2024) as the base model for fine-tuning the Translator, and employs GLM-4.5 (Zeng et al., 2024) for the Planner and Memory modules. We compare CPT-Agent with prompt-based single-LLM baselines with GPT-4o (OpenAI, 2023), GLM-4.5 and Qwen2-7B as the underlying models. For CPT-Agent and each baseline, we initialize one student agent per grade (1-9) and each writing proficiency level (*Low, Medium, High*), resulting in a total of 27 agents to ensure adequate diversity across all experimental conditions. The detailed experimental settings are provided in Appendix E.

5.2 Evaluating Composition Simulation

This experiment evaluates the composition capability of the Translator, with the Memory and Planner of CPT-Agent disabled. Specifically, we use GPT-4o to generate 5 suitable writing prompts for each grade. Agents within the same grade level generate essays based on the same prompts.

LLM-as-Judge Evaluation. Recent studies have shown that DeepSeek-R1 exhibits reliability in writing evaluation (Gao et al., 2025a). Therefore, we employed DeepSeek-R1 as an automated judge to evaluate the generated essays, taking as input the evaluation criteria, the student's grade, and the generated essay, and producing a score (0–100). We computed the *average score* for each proficiency level to verify our core assumption that an agent's generated essays should reflect its predefined writing proficiency.

Human Evaluation. We also conducted a human evaluation. Three annotators, all literature majors at a normal university, were presented with essay triads on the same prompt, where each essay was generated by a simulated student agent at *Low, Medium*, or *High* writing proficiency. The annotators ranked these agents from highest to lowest writing quality based on their essays. We measured *ranking agreement* as the proportion of cases where human rankings exactly matched the predefined proficiency rankings, then computed the *average agreement score* across all annotators.

Results. As shown in Figure 3, CPT-Agent produces a more diverse score distribution across essays generated by simulated students of varying proficiency levels, compared to the baselines, which struggle to generate low-proficiency essays (GPT-4o, GLM-4.5) or to distinguish between levels (Qwen2-7B, GLM-4.5).

Table 1 presents the *average agreement scores* of the compared methods. CPT-Agent achieves 59.26% agreement, substantially outperforming all the baselines. This result highlights a key advantage of fine-tuning: unlike generic LLMs, which tend to produce uniformly fluent text, the fine-tuned Translator can authentically reproduce the characteristic weaknesses of low-proficiency writing.

5.3 Evaluating Revision Simulation

To assess revision simulation, we quantify agents' alignment with expected behaviors across cognitive ability and writing proficiency dimensions. We primarily compare CPT-Agent against the competi-

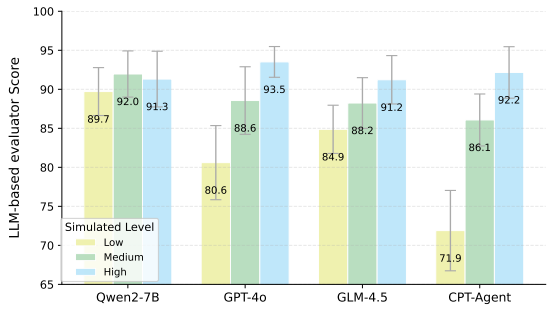


Figure 3: Average scores for essays generated at Low, Medium, and High proficiency levels. CPT-Agent shows clear separation, while baselines struggle with low-proficiency simulation.

Models	Agreement (%)
Qwen2-7B	16.67 ± 4.72
GLM-4.5	31.11 ± 6.67
GPT-4o	42.22 ± 3.85
CPT-Agent	$\mathbf{59.26 \pm 4.14}$

Table 1: Agreement Score: Proportion of cases where human annotators' proficiency rankings exactly match predefined levels. CPT-Agent achieves 59.26%, substantially outperforming all baselines.

tive prompt-based baseline powered by GLM-4.5. For this evaluation, we utilized the synthetic essays generated during the composition simulation evaluation.

5.3.1 Multi-dimensional Feedback Adoption

This evaluation primarily assesses *cognitive filtering* by the Planner and revision quality. We classify writing feedback into six distinct dimensions ranging from surface-level (e.g., error correction) to high-level (e.g., theme development), where higher-level dimensions require more advanced cognitive abilities. For each generated essay, GPT-4o was employed to produce multi-dimensional feedback covering all 6 dimensions (See Appendix E.3).

To evaluate how simulated agents process fine-grained feedback, we conducted human evaluation using *feedback adoption rate* as our metric, which measures the proportion of feedback items within each dimension that were successfully revised.

Results. Figure 4 presents the *feedback adoption rate* of CPT-Agent and the baseline. CPT-Agent demonstrates a developmentally appropriate trajectory. Low-grade students primarily adopt surface-level feedback and progressively incorporate more complex feedback as grade level increases. This progression aligns with established findings in ed-

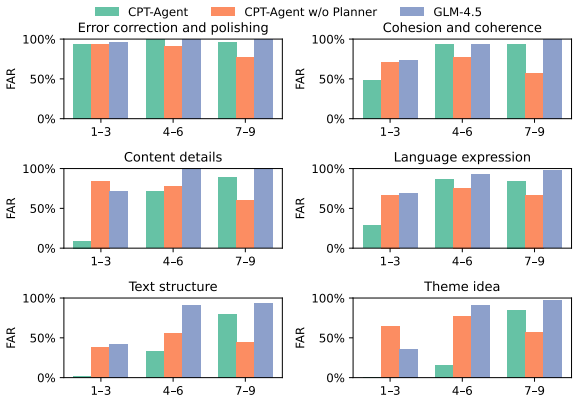


Figure 4: Feedback adoption rate (%) across 6 feedback dimensions. CPT-Agent shows a developmentally appropriate gradient (low-grade students adopt surface feedback first), while the baseline over-applies complex feedback at all levels.

ucational writing research (Piaget, 1971). In contrast, the prompt-based baseline's developmental trajectory is less stable. For example, it achieves a feedback adoption rate of nearly 40% in thematic feedback even for low-grade students, contradicting expected cognitive developmental stages. Additionally, the baseline fails to maintain clearly distinguishable revision behaviors between middle and high grade levels. In addition, removing the Planner yielded results similar to the baseline, though with more constrained feedback adoption. These comparisons demonstrate that the Planner's *cognitive filtering* is essential. Without the cognitive control, LLMs over-apply complex feedback regardless of the student's developmental stage.

5.3.2 Alignment with Instructional Theories

We consider three writing feedback strategies: Process Writing (PW) (Flower and Hayes, 1981), Direct Instruction (DI) (Graham and Perin, 2007), and Self-Regulated Strategy Development (SRSD) (Harris and Graham, 1996), and investigate whether student agents respond to them in ways that are consistent with established theories.

For each generated essay, we employed GPT-4o to provide feedback according to the three specified instructional strategies (Appendix E.4). All original and revised essays were then scored using DeepSeek-R1. We calculated the *relative score changes* between original and revised versions, measuring the improvement attributable to each feedback strategy.

Results. Table 2 shows the *average relative score changes* for student agents across different writ-

Models	CPT-Agent			GLM-4.5		
	PW	DI	SRSD	PW	DI	SRSD
Low	8.80	9.12	17.37	26.96	24.74	30.06
Medium	6.65	5.80	10.91	14.43	9.19	16.71
High	2.23	2.04	5.89	3.72	3.28	7.87
Grade 1-3	2.20	8.21	11.33	14.04	11.71	18.26
Grade 4-6	6.25	4.45	10.44	15.23	11.88	18.70
Grade 7-9	9.14	4.30	12.39	15.84	13.62	17.68

Table 2: Average relative score changes under three feedback strategies. CPT-Agent's patterns match educational theory better.

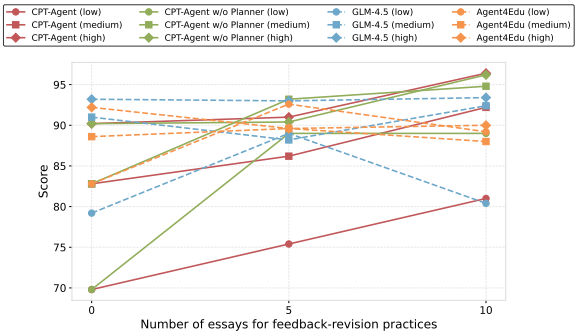


Figure 5: Average scores of agents' newly generated essays after 0, 5, and 10 rounds of feedback-revision practices. CPT-Agent shows gradual improvement, while alternatives degrade (CFR, Agent4Edu) or improve unrealistically fast (no Planner).

ing proficiency levels and grades. We make the following key observations:

(1) *DI vs. PW across grades*: The student agents simulated by CPT-Agent in middle and upper grades show greater improvement with PW feedback compared to DI feedback, whereas the opposite trend is observed among early-primary student agents. These patterns align with well-established educational theories: Students in lower grades are more responsive to the explicit guidance provided by DI, while more advanced learners benefit more from the open-ended, reflective nature of PW feedback (Graham and Perin, 2007).

In contrast, GLM-4.5 demonstrates weaker alignment with these theoretical expectations. For the student agents with both low proficiency and low grade levels, PW feedback is observed to be more effective than DI feedback.

(2) *SRSD advantage*: Across all experimental conditions, SRSD feedback consistently yields the most substantial improvements for both CPT-Agent and baseline agents. This aligns with established research on the efficacy of self-regulated strategy instruction (Harris and Graham, 1996). This suggests that the procedural scaffolding inherent in SRSD mirrors the computational logic of iterative reasoning and reflection in LLMs, providing a structured pathway for the agent to emulate the self-regulatory behaviors of human writers.

(3) *The scale of proficiency improvement*: While all feedback strategies yield score gains across agents, the magnitude of these improvements varies significantly. Notably, the baseline agents exhibit substantial, often unrealistic, score leaps, even for low-proficiency students. This suggests that, in the absence of explicit cognitive constraints, monolithic LLMs tend toward indiscriminate feedback adoption. This failure to decouple cognitive capacity from linguistic capabilities results in over-revision, where the agent may ignore its low-proficiency persona because the underlying LLM is too strong in following instructions and generating fluent text. In contrast, CPT-Agent's modular constraints ensure that growth remains tied to the agent's simulated cognitive boundaries.

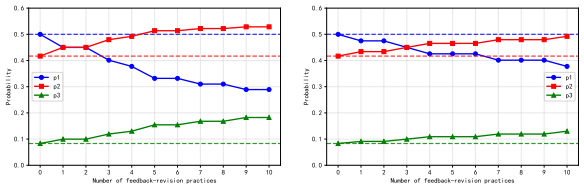
5.4 Writing Proficiency Development

We further evaluate CPT-Agent's ability to learn and apply writing skills over time. Each agent underwent ten rounds of iterative writing and revision, then wrote 5 new essays to assess whether its writing had improved.

We evaluate CPT-Agent against two alternative memory management approaches: (1) Complete Feedback Record (CFR), which uses the full record of past feedback as memory; and (2) Agent4Edu (Gao et al., 2025b), which maintains a sequence of summarized feedback snippets as memory.

Results. To evaluate long-term learning trajectories, we assess essays generated over multiple practice rounds facilitated by DeepSeek-R1. As shown in Figure 5, CPT-Agent demonstrates a stable and progressive improvement in writing quality across practice rounds. In contrast, agents employing unstructured memory records exhibit performance decay, highlighting the importance of our abstracted skill representation in mitigating information interference. Furthermore, the ablation of the Planner leads to disjointed learning spikes. This result highlights the Planner's critical role as a cognitive regulator. Collectively, these experiments demonstrate that structured memory and *cognitive filtering* are both important for pedagogical realism.

Cognitive Level Distribution as a Developmental Monitor. The cognitive level distribution derived from the Ordered Logit Model serves as an internal monitor of cognitive growth, providing a latent de-



(a) To high-quality feedback (b) To sub-optimal feedback

Figure 6: Evolution of the sampled cognitive level distribution for two grade 3 agents across successive feedback-revision cycles.

velopmental metric that complements observable writing performance. Figure 6a illustrates the evolution of these cognitive probabilities for an agent across ten successive cycles of feedback-revision practice. The results indicate that as learning iterations progress, the agent's cognitive state shifts toward higher levels of processing while remaining within the expected developmental bounds of its assigned grade level. To test the framework's sensitivity to feedback quality, we introduced a controlled perturbation by randomly withholding half of the feedback elements for an agent with identical initial settings. As depicted in Figure 6b, sub-optimal feedback attenuates the developmental curve, resulting in a slower rate of cognitive advancement. This shift in probability mass demonstrates the CPT-Agent's capacity to simulate longitudinal cognitive development and the principled internalization of writing strategies. Unlike rigid, deterministic cognitive assignments that overlook the inherent fluidity of learning, our application of the Ordered Logit Model captures the stochastic and transitional nature of student cognition.

5.5 CPT-Agent for Teacher Training

We recruited five novice teachers to evaluate the framework's responsiveness to human feedback. The study utilized a stratified set of six agents (two per grade for grades 2, 5, and 8). Each grade level was paired with an age-appropriate writing prompt to reflect realistic classroom assignments.

Participants provided revision-oriented feedback on the initial essays generated by the agents. To quantify the impact of this feedback, we utilized DeepSeek-R1 to assess essay quality before and after the simulated revisions. For each prompt, *average score gain* across the simulated students served as the primary metric to measure the instructional efficacy of each participant's feedback. This approach allowed for a standardized and scalable evaluation of how effectively different feedback strategies catalyzed student improvement.

Two teaching experts also ranked the novice teachers' feedback quality for each prompt. The average Pearson correlation across prompts between expert rankings and CPT-Agent's automated rankings (based on average score gains) is $r = 0.72$, suggesting promising alignment despite the limited sample size. The prompt-based baseline (GLM-4.5) achieved a correlation of -0.375, as it tends to produce uniformly high-quality revisions regardless of feedback. This result suggests that cognitively grounded student simulation can serve as a proxy for expert evaluation. If simulated students respond to feedback in developmentally realistic ways, their score gains become meaningful indicators of feedback quality.

We also conducted interviews with the participants using five structured questions designed to probe their experience of using CPT-Agent and the baseline. The results show that CPT-Agent consistently achieved higher scores than the baseline in writing, revision simulation and helpfulness in feedback adaptation. Participants found CPT-Agent's cognitive transparency (seeing which feedback was filtered and why) particularly valuable, as it helped them recognize when their feedback exceeded a student's cognitive level. More details of the interviews are provided in Appendix F.3.

6 Conclusion

We presented CPT-Agent, a theoretically grounded framework for simulating student writing behavior that shifts the focus from idealized performance to pedagogical realism. By explicitly decoupling cognitive ability from writing proficiency through a modular design, this work prioritizes behavioral and developmental authenticity and establishes a multidimensional evaluation paradigm to assess agent realism. Our findings demonstrate that while fine-tuning is essential for reproducing the characteristic weaknesses of low-proficiency writing, the implementation of *cognitive filtering* is the critical mechanism for generating realistic, non-linear learning trajectories. By achieving strong alignment with expert judgments in evaluating novice teachers' feedback quality, CPT-Agent provides a robust foundation for both automated feedback assessment and scalable teacher training, proving that the value of a simulated student lies in its principled limitations that mirror authentic human learning.

7 Limitations

Our work has several limitations that suggest valuable directions for future research.

First, while CPT-Agent is grounded in the Cognitive Process Theory of Writing, it represents a stylized abstraction of human cognition. The framework primarily focuses on the linear flow between planning, translating, and knowledge accumulation. However, actual human writing involves highly non-linear, recursive, and subconscious interactions, such as the emotional state of the learner. Such factors are currently beyond the scope of our modular architecture. Future work could explore more dynamic interaction loops to better capture the affective dimensions of the writing process.

Second, our current experiments focus on teacher-to-student text-based feedback. While this is a critical pedagogical scenario, it does not account for the multi-modal and social nature of writing development. In real-world environments, students benefit from peer-to-peer reviews, collaborative group discussions, and multi-modal feedback (e.g., audio or visual annotations). The current framework is optimized for text-based instructional feedback and may require further adaptation to simulate the social-constructivist dynamics of a collaborative writing environment.

Third, while we simulate memory and growth across multiple cycles, these simulations occur within a relatively short experimental window. In reality, student writing development is a longitudinal process spanning years, influenced by cumulative background knowledge and maturing cognitive faculties. Our model approximates this growth through parameter shifts, but long-term empirical studies are needed to verify if the agent's learning curve truly mirrors the multi-year trajectory of a human student.

Finally, although our evaluation involved expert educators and novice teachers, the sample size remains relatively small and localized. Expanding the evaluation to include a broader demographic of both educators and students from various regions would provide a more robust validation of the framework's applicability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62376166, 62306188) and the Academy for Multidisciplinary Studies of Capital Normal University.

References

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using llms to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368.

Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined ai personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Saptarshi Bhowmik, Luke West, Alex Barrett, Nuodi Zhang, Chih-Pu Dai, Zlatko Sokolikj, Sherry Southerland, Xin Yuan, and Fengfeng Ke. 2024. Evaluation of an llm-powered student agent for teacher training. In *European conference on technology enhanced learning*, pages 68–74. Springer.

Ngoc My Bui and Jessie S Barrot. 2025. Chatgpt as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, 30(2):2041–2058.

Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2024. Empowering private tutoring by chaining large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 354–364.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025. Llm agents for education: Advances and applications. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 13782–13810. Association for Computational Linguistics.

Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, Christopher Bogart, Eric Keylor, Can Kültür, Jaromír Savelka, and Majd Sakr. 2024. A comparative study of ai-generated (GPT-4) and human-crafted mcqs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference, ACE 2024, Sydney, NSW, Australia, 29 January 2024- 2 February 2024*, pages 114–123. ACM.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication*, 32(4):365–387.

Huixin Gao, Harwati Hashim, and Melor Md Yunus. 2025a. Assessing the reliability and relevance of deepseek in efl writing evaluation: a generalizability theory approach. *Language Testing in Asia*, 15(1):33.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang.

2025b. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23923–23932.

Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A design space for writing support tools using a cognitive process model of writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 11–24, Dublin, Ireland. Association for Computational Linguistics.

Steve Graham, Karen Harris, and Michael Hebert. 2011. Informing writing: The benefits of formative assessment. a report from carnegie corporation of new york. *Carnegie Corporation of New York*.

Steve Graham and Dolores Perin. 2007. Writing next-effective strategies to improve writing of adolescents in middle and high schools.

Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and Alice Oh. 2023. RECIPE: how to integrate chatgpt into EFL writing education. In *Proceedings of the Tenth ACM Conference on Learning @ Scale, Copenhagen, Denmark, July 20-22, 2023*, pages 416–420. ACM.

Karen R Harris and Steven Graham. 1996. *Making the writing process work: Strategies for composition and self-regulation*. Brookline Books, Cambridge, Mass.

John R Hayes. 2012. Modeling and remodeling writing. *Written communication*, 29(3):369–388.

Ken Hyland and Fiona Hyland. 2019. *Feedback in second language writing: Contexts and issues*. Cambridge university press.

Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach ai how to code: Using large language models as teachable agents for programming education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Breanna Jury, Angela Lorusso, Juho Leinonen, Paul Denny, and Andrew Luxton-Reilly. 2024. Evaluating llm-generated worked examples in an introductory programming course. In *Proceedings of the 26th Australasian computing education conference*, pages 77–86.

Haoxuan Li, Jifan Yu, Xin Cong, Yang Dang, Daniel Zhang-li, Yisi Zhan, Huiqin Liu, and Zhiyuan Liu. 2025. Exploring llm-based student simulation for metacognitive cultivation. *arXiv preprint arXiv:2502.11678*, arXiv:2502.11678.

Yu Li, Shang Qu, Jili Shen, Shangchao Min, and Zhou Yu. 2024. Curriculum-driven edubot: A framework for developing language learning chatbots through synthesizing conversational data. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 400–419.

Zhengyuan Liu, Stella Yin, Geyu Lin, and Nancy Chen. 2024. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642.

Zhexiong Liu, Diane Litman, Elaine L Wang, Tianwen Li, Mason Gobat, Lindsay Clare Matsumura, and Richard Correnti. 2025. erevise+ rf: A writing evaluation system for assessing student essay revisions and providing formative feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 173–190.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Irum Naz and Rodney Robertson. 2024. Exploring the feasibility and efficacy of chatgpt3 for personalized feedback in teaching. *Electronic Journal of e-Learning*, 22(2):98–111.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. 2025. Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Jean Piaget. 1971. The theory of stages in cognitive development. In D. R. Green, M. P. Ford, and G. B. Flamer, editors, *Measurement and Piaget*. McGraw-Hill.

Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. Large language models for education: Grading open-ended questions using chatgpt. In *Simpósio Brasileiro de Engenharia de Software (SBES)*, pages 293–302. SBC.

Robert S Siegler. 1998. *Emerging minds: The process of change in children's thinking*. Oxford University Press.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. Class: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the*

Association for Computational Linguistics: EMNLP 2023, pages 1941–1961.

KV Srivatsa, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2025. Can llms reliably simulate real students' abilities in mathematics and reading comprehension? *arXiv preprint arXiv:2507.08232*.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. 2025. A cognitive writing perspective for constrained long-form text generation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 9832–9844. Association for Computational Linguistics.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2025a. Large language models for education: A survey and outlook. *IEEE Signal Process. Mag.*, 42(6):51–63.

Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. 2025b. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 510–519.

Songlin Xu and Xinyu Zhang. 2023. Leveraging generative artificial intelligence to simulate student learning behavior. *arXiv preprint arXiv:2310.19206*, arXiv:2310.19206.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning. *arXiv preprint arXiv:2404.07963*, arXiv:2404.07963.

Yu Yuan, Lili Zhao, Wei Chen, Guangting Zheng, Kai Zhang, Mengdi Zhang, and Qi Liu. 2025. Simulating human-like learning dynamics with llm-empowered agents. *arXiv preprint arXiv:2508.05622*.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, and 36 others. 2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965.

Guanhao Zhao, Zhenya Huang, Yan Zhuang, Jiayu Liu, Qi Liu, Zhiding Liu, Jinze Wu, and Enhong Chen. 2023. Simulating student interactions with two-stage imitation learning for intelligent educational systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3423–3432.

A Theoretical Motivation

The Cognitive Process Theory of Writing (Flower and Hayes, 1981) conceptualizes writing as a goal-directed, recursive activity involving three core components: The writer's long-term memory, the writing process, and the task environment. The long-term memory includes knowledge of topic, audience, and writing conventions. The writing process encompasses planning, translating, and reviewing. Planning involves building an internal representation of knowledge to guide writing; translating refers to converting this representation into written language; and reviewing entails evaluating and revising the text. The task environment refers to contextual factors present during the writing task, including the writing prompt and feedback.

This theoretical framework is well suited for simulating authentic student writing behavior, particularly the critical processes of feedback integration and revision. While recent studies have applied this theory to design writing assistance tools (Gero et al., 2022; Wan et al., 2025), we pioneer its application to simulating student writing behavior and evaluating feedback effectiveness.

We note that the revised CPT model (Hayes, 2012) also incorporates motivation, affect, and audience awareness; the current work focuses on the core cognitive-procedural components and leaves these extensions for future exploration.

B Planner Configuration

B.1 Cognitive Ability Profile

The cognitive ability to respond to feedback is categorized into three developmental levels correspond-

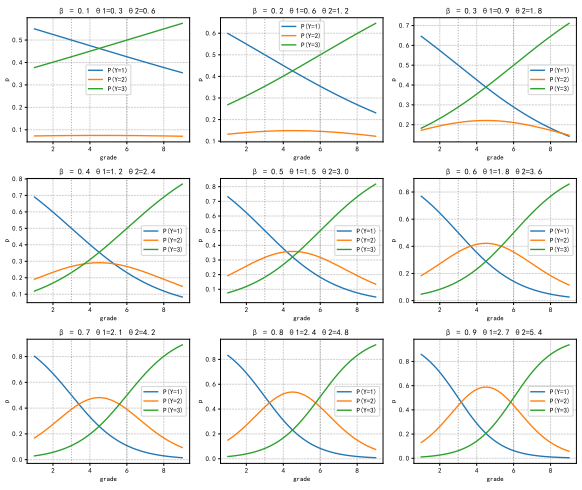


Figure 7: Curve with β ranging from 0.1 to 0.9

ing to grade levels: *Low, Moderate*, and *Strong*.

Low (Grades 1–3):

- Low feedback comprehension ability: Can only process surface-level issues such as word choice, punctuation, and sentence fluency.
- Feedback needs: Requires highly specific, immediate, and narrowly focused feedback (ideally accompanied by concrete examples).

Moderate (Grades 4–6):

- Moderate feedback comprehension ability: Able to understand issues such as content repetition, overly simple sentence structures, and unclear openings or conclusions.
- Feedback needs: Needs clear prompts on aspects such as local structure, use of transitional words, and language richness.

Strong (Grades 7–9):

- Strong feedback comprehension ability: Capable of understanding complex feedback, such as insufficient argumentation, unclear structure, or vague central ideas.
- Feedback needs: Best suited for receiving developmental feedback, such as guidance on writing strategies and suggestions for self-improvement.

B.2 Settings of the Ordered Logit Model

Through preliminary testing, we find that when beta >1, the values on both sides are zero. Therefore, we initially set the beta value in the range of 0-1. As shown in Figure 7, We plotted graphs for β values from 0.1 to 0.9 and observed that β values between 0.7 and 0.9 produced reasonable performance. To effectively simulate the cognitive comprehension abilities of students across different grade levels, we broadly categorized grades 1–3, grades 4–6, and grades 7–9 into distinct cognitive levels. And the parameters were set as follows: $\theta_1 = 2.4, \theta_2 = 4.8$, and $\beta = 0.8$.

The parameter α_i represents a score between 0 and 1, indicating an estimation of the extra cognitive capability based on student's writing skills. An LLM is used to estimate this score. The instruction for the LLM is shown in Table 3.

 # Personalized Cognitive Parameter Assessment
- **Target Parameter:** α_i (Extra Cognitive Ability)
- **Cognitive Level Definitions:** {Cognitive Level Descriptions}
- **Current Writing Skills (K):** {Writing Skills}

 # Task: Psychometric Estimation
Evaluate the student's mastery of writing skills K relative to the provided Cognitive Level Definitions. Your goal is to map the qualitative evidence of their skills onto a continuous quantitative scale [0, 1].

 # Evaluation Rubric

- **Low:** 0.0 – 0.3
- **Moderate:** 0.4 – 0.7
- **Strong:** 0.8 – 1.0

 # Constraints
- Provide the score as a single floating-point number between 0 and 1.
- Ensure the score is strictly derived from the intersection of K and the Cognitive Levels.
- **Output Format:** [Score] (No text, no preamble).

Table 3: The instruction for estimating the personalized cognitive parameter α_i.

B.3 The Prompt for the Planner

The prompt for CPT-Agent's Planner is shown in Table 4.

C Memory Configuration

C.1 Hierarchical Skill Knowledge base

We design a tree-structured knowledge base to organize writing skills (defined as writing tips or strategies) as a form of long-term memory. The hierarchical structure is organized into four top-level categories:

- Basic Skills
  - Topic and Content

 # Profile Configuration
You are a student in {*Student Grade*}.
Cognitive Levels: {*Cognitive Level Descriptions*}
Current Cognitive State (Distribution): $\{\{P(Y = j)\}_{j=1}^{3}\}$
Act as a cognitive filter: Process information strictly according to the probability distribution of your developmental stages.

 # Task: Cognitive Processing
1. Filtering Strategy:

- *Scenario A (Initial Prompt):* If teacher feedback is empty, analyze the writing prompt {*prompt*}. If your internal writing knowledge is insufficient, output the prompt verbatim. Otherwise, adapt your knowledge into specific writing intentions.
- *Scenario B (Revision):* If feedback {*feedback*} is provided, extract only the suggestions that align with your dominant cognitive levels. Discard suggestions that exceed your current comprehension threshold.

2. Instruction Generation: Transform the filtered suggestions into a concise, internal writing plan.

 # Constraints
- **Zero-Hallucination:** Do not include suggestions that were not in the original feedback or your retrieved knowledge.
- **Format:** Output *only* the [Writing Plan].
- **Brevity:** Total output must not exceed 200 characters. No meta-commentary.

Table 4: The prompt template for the CPT-Agent's Planner component.

  - Vocabulary and Grammar
  - Rhetoric and Style
  - Emotional Expression
- Narration Skills
- Argumentation Skills
- Expository Skills

The hierarchy serves as an initial structure that captures the major dimensions of writing development. It can be extended by adding sub-nodes under the second-level categories to represent more fine-grained writing skills.

C.2 The Prompt for Knowledge Integrator

The prompt for CPT-Agent's Knowledge Integrator is shown in Table 5.

D Translator Configuration

D.1 Composition Proficiency Profile

Student foundational composition proficiency is defined across three levels: *Low, Medium*, and *High* (Graham et al., 2011), with specific descriptions as follows:

Low: Perform below the average level of their peers in writing tasks, exhibiting significant difficulties in content development, organization, language expression, or writing fluency. Common characteristics include:

- Content: Unclear arguments and a lack of supporting details or examples.
- Organization: Poor structural coherence and weak logical connections between paragraphs.
- Language: Frequent grammatical errors, limited vocabulary, and monotonous sentence structures.
- Behavioral: Low writing motivation and weak

 # Role Context
You are a Knowledge Integration module responsible for student skill acquisition. Your task is to update the student's *writing skill base* by synthesizing recent comprehensible writing suggestions into generalizable procedural knowledge.
- **Current Cognitive Profile:** {*Cognitive Profile*}

 # Input Data
Suggestions (new knowledge): {*Suggestions*}
Current writing skill base (existing knowledge): {*Skills*}

 # Task Instructions
1. Extraction: Identify pedagogical strategies from the suggestions that are generalizable across different writing contexts.
2. Consolidation: Compare new findings with the Existing State. If a new strategy is similar to an existing one, merge them. If it is novel, add it to the appropriate category.
3. Category Mapping: Every strategy must be mapped to one of the following predefined categories: {Category Hierarchy}.

 # Constraints
- **Evidence-Based:** Do not invent strategies. Every update must be derived from the provided Dialogue History.
- **Generalizability:** Focus on how-to procedural knowledge rather than specific corrections for a single essay.
- **Preservation:** Retain all existing strategies that are not contradicted or superseded by new evidence.
- **Cognitively-Aligned:** Use cognitively appropriate language style to describe the skills.

 # Output Format
1. Output a valid JSON object only. Use the structure: `{"Category": {"Strategy_Name": "Procedural content"}}`
2. Follow the JSON with a *Revision Summary*: A one-sentence technical description of which categories were updated.

Table 5: The prompt template for the CPT-Agent's Knowledge Integrator, focusing on longitudinal skill acquisition and memory consolidation.

self-regulation abilities (e.g., difficulty setting goals or self-assessing), often accompanied by writing anxiety.

Medium: Writing performance is close to the average level of their peers. They are able to complete basic writing tasks but may show room for improvement in complexity, depth, or stylistic expression. Common characteristics include:

- Content: Able to present basic arguments, but supporting details may be limited or lack persuasiveness.
- Organization: The overall structure is generally clear, though transitions between paragraphs or the logical flow of the essay may be somewhat weak.
- Language: Few grammatical errors; vocabulary use is appropriate but lacks variety; sentence structures meet basic requirements.
- Behavioral: Demonstrates some self-regulation (e.g., can set simple goals), but may lack advanced strategies such as rhetorical refinement.

High: Perform significantly above the average level of their peers in writing tasks, demonstrating strong creativity, logical reasoning, and expressive language skills. Common characteristics include:

- Content: Arguments are clear and insightful, with rich and persuasive supporting details.
- Organization: The essay exhibits a well-structured and coherent organization, with smooth paragraph transitions and a logically sound overall layout.
- Language: Grammar is accurate; vocabulary is rich and varied; sentence structures are complex and flexible.
- Behavioral: Strong self-regulation skills; capable of setting and pursuing complex goals (e.g., enhancing reader engagement); shows critical thinking during the revision process.

D.2 Revision Proficiency Profile

- **Low**: Students struggle to implement feedback correctly even when comprehended, often making incomplete or erroneous revisions.
- **Medium**: Students can apply feedback with clear guidance but lack independent revision capabilities.
- **High**: Students demonstrate self-regulated revision by reflecting on feedback and autonomously improving their writing.

D.3 The Prompt for the Translator

The prompt for CPT-Agent's Translator is shown in Table 6.

 # Role Context
You are a student in {*Student Grade*}. Your writing behavior must strictly adhere to the following profiles:
- **Composition Proficiency:** {*Composition Proficiency Level*}
- **Revision Proficiency:** {*Revision Proficiency Level*}
Do not exceed these capabilities; maintain the linguistic patterns and cognitive limitations typical of this profile.

 # Task Instructions
1.Analysis: Evaluate the provided input to determine if it is a new writing prompt or revision feedback.
2. Execution:
• *If a prompt:* Compose an initial essay that reflects your assigned grade and proficiency.
• *If revision feedback:* Update the existing essay. Your revisions must only reflect the improvements you are capable of, based on your *Feedback Application Proficiency*.

 # Constraints
- Maintain consistency in tone and error patterns (if applicable to the level).
- Output only the resulting essay text.

Table 6: The prompt template for the CPT-Agent's Translator component.

D.4 Fine-tuning Details

D.4.1 Fine-tuning

We aim to employ a smaller open-source model (Qwen2-Instruct-7B) fine-tuned to generate writing outputs that reflect varying levels of proficiency, while preserving the model's robust instruction-following capabilities. To this end, we curate an essay dataset for fine-tuning. The data is a set of essays collected from the Web, which contains writing samples from students ranging from Grade 1 to Grade 9. Each sample includes a title, content, rating (ranging from 1 to 5) and grade level.

Writing Data: We obtained the essay dataset with permission for research use. The original writing data contains information such as grade level, essay topic, essay category, score, and evaluation. After removing entries containing personal information, we further processed the dataset. Based on the rating and their distribution, we categorized

essays into three proficiency levels: *Low* (ratings of 1 or 2), *Medium* (rating of 3), and *High* (ratings of 4 or 5). For each grade, we then sample 1000 essays per writing proficiency level to construct a dataset. We employed several high school teachers to review the sampled data and removed any essays whose quality clearly did not align with their assigned ratings.

Revision Data: To ensure robust instruction-following capability during revisions, we employ knowledge distillation from GPT-4o for dataset generation. The distillation pipeline comprises three stages:

1. **Target Sampling**: For different writing proficiency levels, select representative essays from the essay dataset as revision targets.

2. **Feedback Generation**: GPT-4o acts as an instructor to generate feedback based on the provided essay content.

3. **Revision Simulation**: GPT-4o simulates a corresponding level and grade student and producing revised essays based on generated feedback.

We observe that GPT-4o could provide comprehensive feedback, but the simulated students frequently produce overly polished revisions (Naz and Robertson, 2024). To address this, we compute cosine similarity based on BERT-Score (Zhang et al., 2019) between original and revised essays, heuristically excluding samples with similarity < 0.85. The combined dataset is used for fine-tuning the Translator.

D.4.2 Fine-tuning Data Format

We combined both datasets to fine-tune the Translator module.

Data for Diverse Levels of Writing Proficiency

Instruction: Student profile, task description

Input: Writing prompt

Output: Initial essay

The data is used to train the Translator to generate an initial draft based on the student's grade level, composition proficiency and a given writing prompt.

Data for Revision Instruction Following

Instruction: Student profile, task description, writing prompt

Input: Initial draft and instructor's feedback (revision suggestions)

Output: Revised essay

The data is used to train the model to revise an existing draft based on teacher feedback.

D.4.3 Fine-tuning Parameter and Computing Infrastructure

The base model is Qwen2-7B-Instruct model. The fine-tuning configurations for student writing models are detailed in Table 7 and Table 8. All fine-tuning experiments were conducted on machines equipped with NVIDIA A100-PCIE-40GB GPUs, running Ubuntu 22.04.3 LTS (GNU/Linux 5.15.0-141-generic x86_64). The models were implemented using the PyTorch framework, with CUDA 12.1 and PyTorch version 2.6.0.

Parameter	Value
data	8635
learning_rate	3e-05
train_batch_size	4
eval_batch_size	4
seed	42
total_train_batch_size	32
lr_scheduler_type	cosine
num_epochs	2.0
Lora rank	8

Table 7: Parameters for first-stage fine-tuning the Translator.

Parameter	Value
data	8635
learning_rate	1e-05
train_batch_size	4
eval_batch_size	4
seed	42
total_train_batch_size	32
lr_scheduler_type	cosine
num_epochs	3.0
Lora rank	12

Table 8: Parameters for second-stage fine-tuning the Translator.

We applied 4-bit quantization to the fine-tuned Qwen model, reducing GPU memory usage while preserving the model's performance. Experimental results in Section 5.2 demonstrate that the quantized Translator maintains its core capabilities. Consequently, in all subsequent experiments, we employed the quantized model as the underlying architecture for the Translator.

D.4.4 The Prompt for LLM-based Instructor

The prompt for LLM-based Instructor is shown in Table 9.

 # Role Context
You are an expert language educator specializing in composition pedagogy. Your objective is to provide actionable, scaffolded feedback that facilitates writing development.

 # Student Context
Target Profile: {*Student Profile*}

 # Principles

- **Differentiated Instruction:** Align the complexity of your critique with the student's current proficiency.
- **Heuristic Guidance:** Encourage student agency and guide them to learn how to analyze problems.
- **Sequential Progression:** Teach step by step, according to the student's cognitive development.

 # Task: Diagnostic Feedback
Evaluate the student's response to the prompt: {*Writing Prompt*}. Provide prioritized, specific, and constructive advice.

 # Output Requirements
- Maintain a supportive yet rigorous professional tone.
- Do not provide a rewritten version of the essay.
- Format the response strictly as:
[Revision Suggestions]

Table 9: The prompt template for the LLM-based Instructor for collecting data to fine-tuning the Translator.

D.4.5 The Prompt for LLM-based Student Agent

The prompt for LLM-based Student Agent is shown in Table 10.

 # Role Context
You are a student in {*Student Grade*}.
Composition Proficiency: {*Composition Proficiency Profile*}

 # Task
Your goal is to revise your essay. Read the provided revision feedback and revise your essay accordingly. You must remain strictly within the linguistic and cognitive boundaries of your profile. Do not exhibit vocabulary, sentence complexity, or structural logic that exceeds your assigned level.
 # Constraint
Do not fix errors that are characteristic of your proficiency level unless specifically instructed to do so.

 # Output Requirements
- Provide the complete revised essay only.

Table 10: The prompt template for the LLM-based Student Agent for collecting data for fine-tuning the Translator.

E Experimental Settings

E.1 Baselines

We establish prompt-based single-LLM baselines by instantiating student agents through LLMs (GPT-4o, GLM-4.5, or Qwen2-7B) initialized with structured profiles. Each profile contains: (1) Target grade level, (2) Predefined writing proficiency level, feedback application proficiency, and (3) Cognitive ability aligned with developmental stages, identical to the configuration used in CPT-Agent. In this baseline approach, the simulated agent performs writing and revision tasks through black-box inference, relying on the LLM to implicitly synthesize the configured profile attributes when responding to prompts and feedback.

E.1.1 The Prompt for the Baselines

The prompt for the baselines is shown in Table 11.

E.2 The Prompt for DeepSeek-R1 Essay Scorer

The instruction for DeepSeek-R1-based essay scorer is shown in Table 12.

E.3 The Prompt for GPT-4o to Produce Multi-dimensional Feedback covering Six Categories

The prompts for GPT-4o to produce multi-dimensional feedback covering all six categories is shown in Table 13.

E.4 The Prompt for GPT-4o to Generate Three Types of Feedback

The prompts for GPT-4o to generate three types of feedback is shown in Table 14.

F More Experimental Results

F.1 Evaluating Composition Simulation

As shown in Figure 8, we present the confusion matrices for the four models. Notably, CPT-Agent achieves higher agreement score for low-proficiency writing data compared to medium and high levels, indicating its effectiveness in simulating the struggles of low-proficiency students and partially alleviating the challenge faced by LLMs in capturing the struggles and confusion that characterize real low-performing students.

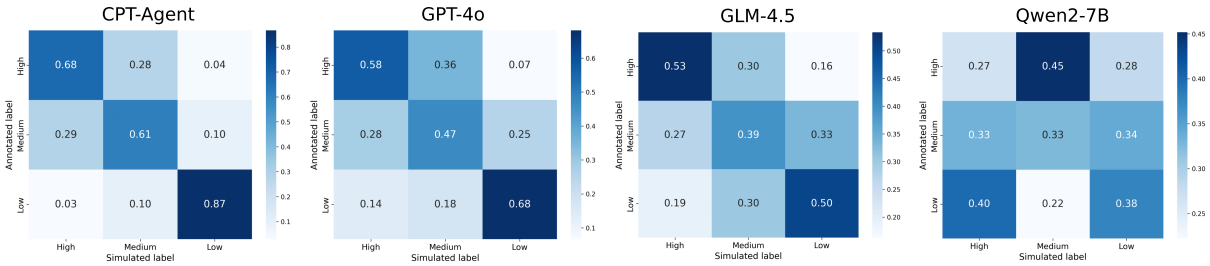


Figure 8: Confusion Matrices of Agreement Score for the CPT-Agent, GPT-4o, GLM-4.5 and Qwen2-7B.

 # Role Context
Grade Level: {*Student Grade*}
Basic Writing Proficiency: {*Writing Proficiency Profile*}
Cognitive Ability: {*Cognitive Ability Profile*}
Feedback Application Skill: {*Feedback Application Proficiency*}
Constraint: You must strictly embody this persona. Your output must reflect the vocabulary, sentence structure, and cognitive limitations inherent to this specific profile.

 # Task Instructions
1. Analysis & Filtering: Review the input (Prompt or Feedback). Based on your *Cognitive Ability*, identify only the information you can realistically comprehend. Discard any suggestions that are too abstract or advanced for your profile.
2. Strategic Planning: Formulate a revision plan. If the input is a prompt, plan an initial draft. If it is feedback, transform the filtered suggestions into a concise, internal writing plan using your *Feedback Application Skill*.
3. Execution (Writing/Revision): Produce the final essay. Ensure the writing style is consistent with your *Writing Proficiency* and does not accidentally exceed your grade level.

 # Output Requirements
- **Content:** Provide the complete essay text only.
- **Style:** Maintain realistic student-level errors or simplicity as defined by your profile.

Table 11: The baseline prompt template, integrating proficiency and cognitive parameters into a single-stage execution for performance comparison.

 # Role Context:
You are a professional essay scorer.
 # Task Instructions:
The total score is 100 points. I will provide you with the scoring criteria, the student grade level, the writing prompt, and essays on the same topic. You must use a consistent standard to evaluate all essays and assign a score to each one.
Scoring Criteria (Total: 100 points):
Content and Message (40 points): Clear theme, rich content, positive values, and genuine emotion.
Structure and Logic (30 points): Complete structure, clear paragraphing, smooth transitions, and logical coherence.
Language and Expression (30 points): Fluent sentences, accurate word choice, appropriate use of rhetorical devices, and vivid language.
 # Input:
Student Grade Level:{*Student Grade*}
Writing Prompt:{*Writing Prompts*}
Essay Content:{*Writing Data*}
 # Output:
Specific scores and brief evaluations.

Table 12: The prompt for DeepSeek-R1-based rating expert.

F.2 Evaluating Revision Simulation

The Heatmap of relative score changes of *Low*, *Medium*, and *High* Proficiency simulated agents by three feedback strategies is shown in Figure 9.

F.3 CPT-Agent for Teacher Training

We further investigate whether CPT-Agent can assist in evaluating real teacher feedback. We let CPT-Agent serve as a proxy evaluator for assessing the quality of novice teachers' feedback.

F.3.1 Participants

We recruited five undergraduate students majoring in language education from the college of education at a normal university, assigning them the role of novice teacher candidates, with a remuneration of 50 RMB per hour.

For the simulation, we selected 6 agents based on CPT-Agent as representative student agents, two each from grades 2, 5, and 8. Each agent generated

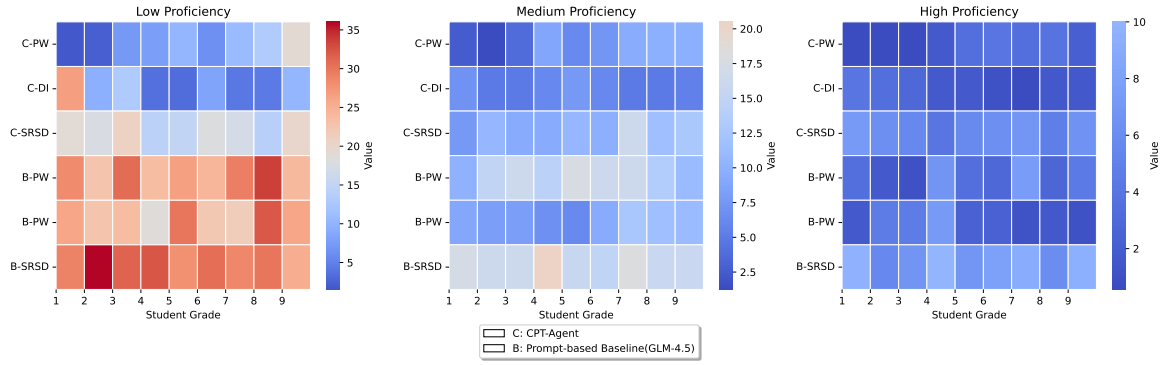


Figure 9: Heatmap of relative score changes of Low, Medium, and High Proficiency simulated agents by three feedback strategies (PW, DI, SRSD).

an essay based on a given writing prompt. The participants then independently interacted with the agents, providing revision support as they would to real students. For the stability of the experimental results, we initialized a essay prompt for each agent. This process yielded 60 sets of conversation data, derived from 2 methods, 5 participants, and 6 agents. The same configuration was applied to the baseline model.

F.3.2 Evaluation

We used DeepSeek-R1 as the scorer to measure changes in essay scores, which are used to reflect the effectiveness of the feedback provided by novice teachers. In addition, we recruited two high school teaching experts to assess the performance of novice teachers based on the feedback they provided. For each agent interaction workflow, the expert needs to rank the novice teachers' ability. For experts, the remuneration is set at 80 RMB per hour.

F.3.3 Questionnaire

Furthermore, we conducted interviews with them using five structured questions designed to probe novice teachers' experience. Each response was rated on a 5-point scale. The scoring criteria is shown in Table 15.

To evaluate the cognitive authenticity and practical utility of the simulated agents, five participants assessed both CPT-Agent and the GLM-4.5 baseline.

As shown in Table 16, CPT-Agent consistently achieved higher average scores across the first three questions (3.17, 2.98, 3.15 against 2.52, 1.93, 1.71), indicating stronger perceived credibility. Especially, the Q3 score of 3.15 indicates that CPT-Agent is effective in helping novice teachers practice and adapt their writing feedback to student needs.

The evaluation of Q4 and Q5 focused on the specific utility of the Planner and Memory modules. The score of 3.47 for Q4 specifically confirms that the simulated students can effectively translate the Planner's instructions into coherent essay revisions. Furthermore, participants reported that the filtered revision suggestions from the Planner were instrumental in revising their own feedback, as it highlighted their prior oversight of student cognitive load management.

Q5, concerning the Memory module's knowledge base, received a score of 3.10. This was a result that combined the student's cognitive profile and the quality of the skills. Participants also provided constructive feedback on the knowledge representation, emphasizing that writing skills should be encapsulated as a balance of concrete, actionable tips and general strategies, rather than leaning too heavily into abstract generalizations. Participants also provided constructive feedback on the knowledge representation, emphasizing that writing skills should be encapsulated as a balance of concrete, actionable tips and general strategies, rather than leaning too heavily into abstract generalizations.

 # Role Context:
Please remember your character identity. All your responses must be based on your identity: You are an experienced middle school Chinese language teacher, skilled in grading essays and proficient in all knowledge related to essay evaluation at the primary and secondary school levels.
Principles:
You must adhere to the following teaching principles:
1. Principle of Individualized Instruction: Provide differentiated teaching based on students' varying abilities and needs.
2. Heuristic Teaching: Stimulate students' initiative and guide them to learn how to analyze problems.
3. Step-by-Step Progression: Teach in a gradual and systematic manner to ensure solid progress.
 # Task Description:
You are revising a student's essay. You need to evaluate the essay based on the writing requirements and provide revision suggestions for the following six aspects:
- Error correction and polishing
- Clarity and coherence
- Content details
- Language expression
- Essay structure
- Thematic depth

Provide one revision suggestion for each aspect.
 # Input:
- Student Grade
- Writing Prompt
- Essay content

 # Output Format:
Error correction and polishing: [Suggestion]
Clarity and coherence: [Suggestion]
Content details: [Suggestion]
Language expression: [Suggestion]
Essay structure: [Suggestion]
Thematic depth: [Suggestion]

Table 13: The prompt for GPT-4o to Produce Multi-dimensional Feedback Covering six Categories.

 # Definition:
These three types of feedback (Process Feedback, Direct Feedback, and Self-Regulated Strategy Development (SRSD) Feedback) are common strategies used in writing instruction and formative writing evaluation. Each focuses on different aspects of writing and supports students in different ways:
1. Process Feedback Definition: Process feedback refers to ongoing feedback provided by teachers or systems throughout various stages of the writing process, such as brainstorming, drafting, revising, and editing. This type of feedback emphasizes cognitive processes and writing strategies rather than evaluating only the final product. Characteristics: Emphasizes writing as a multi-stage process rather than a one-time output. Feedback may address topic selection, idea development, organization, language expression, rhetorical devices, or revision directions. Encourages students to reflect on their writing strategies and gradually develop good writing habits.
2. Direct Feedback Definition: Direct feedback involves the teacher explicitly pointing out specific problems in the student's writing and providing direct suggestions or correct answers. Characteristics: Clear and efficient, suitable for correcting lower-level issues such as grammar, spelling, punctuation, or word usage. Typically provided after the first draft or final version. May not promote deeper cognitive engagement or strategy development.
3. Self-Regulated Strategy Development (SRSD) Feedback Definition: SRSD feedback is grounded in self-regulated learning theory and aims to help students develop writing strategies and self-management skills. It is a form of instructional feedback that not only tells students what to write, but also how to write and how to monitor their writing behavior. Characteristics: Provides explicit instruction in writing strategies (e.g., planning, organizing, revising). Emphasizes self-monitoring, self-evaluation, and goal setting. Often includes modeling, guided practice, and gradual release of responsibility (scaffolding).
 # Input:
- Student Grade
- Writing Prompt
- Essay content

 # Task Description:
I will now provide you with an essay. Please generate three types of feedback, each in the form of a single paragraph.

Table 14: The prompt for GPT-4o to Generate Three Types of Feedback.

Question	Scoring Criteria
Q1: Do the simulated students' revision behaviors (e.g., the adoption of revision suggestions and manner of their edits) align with their designated age and cognitive ability level?	1 - Severe Mismatch: Revision behaviors are developmentally inappropriate and fundamentally misaligned with the student's cognitive abilities. 2 - Partial Alignment: Revision behaviors are occasionally suitable but the adoption of revision suggestions contain noticeable inconsistencies with the student's cognitive level. 3 - General Alignment: Revision behaviors are broadly appropriate and the adoption of revision suggestions generally correspond to the student's cognitive abilities, with minor discrepancies. 4 - High Alignment: Revision behaviors are well-tailored and the adoption of revision suggestions consistently support the student's cognitive development stage. 5 - Perfect Alignment: Revision behaviors are optimally challenging and supportive, demonstrating a nuanced understanding of the student's cognitive potential.
Q2: Does the writing content produced by the simulated students (e.g., word choice, examples used, application of knowledge) consistently match their simulated background profile, such as their grade level and writing proficiency levels?	1 - Severely Inconsistent: The content demonstrates a fundamental misalignment with the core facts or setting of the simulated background. 2 - Frequently Inconsistent: The writing regularly introduces elements that contradict or fall outside the simulated background. 3 - Generally Consistent: The core narrative and intent are in line with the background, though there may be minor inaccuracies in style or detail. 4 - Highly Consistent: The writing is seamlessly aligned with the background in tone, knowledge, and perspective, demonstrating a clear understanding. 5 - Perfectly Consistent: The writing is not only flawless in its alignment but also demonstrates a nuanced understanding, offering unique insights that enrich the simulated persona.
Q3: Does interacting with the simulated students effectively help you practice how to tailor and adapt your feedback to a learner's specific developmental stage?	1 - Ineffective: The interaction did not help me practice tailoring feedback. The students did not exhibit distinct developmental stages to adapt to. 2 - Slightly Effective: The interaction provided minimal opportunity for practice. I could detect slight differences but could not meaningfully adapt my approach. 3 - Moderately Effective: The interaction helped me practice basic adaptation. I could adjust my feedback in a general way to broad developmental differences. 4 - Very Effective: The interaction was largely successful in helping me practice. I could consistently tailor my feedback to specific developmental stages with some precision. 5 - Highly Effective: The interaction was exceptionally valuable for practice. It forced me to refine my feedback strategically and deepened my understanding of developmental adaptation.
Q4: Do the simulated students' revisions accurately execute the instructions provided by the Planner?	1 - Fails to execute. The modified text is irrelevant to the instructions or completely undermines the original text. 2 - Partially executes the instructions. 3 - Executes most of the instructions. 4 - Accurately and fluently executes the vast majority of instructions. 5 - Perfectly executes all instructions.
Q5: Can you observe the student's knowledge base being updated and utilized in a logical way and in a way that is consistent with their profile as they process your feedback?	1 - Ineffective No observable evidence that the student updates or utilizes their knowledge base while processing feedback. The student's responses show no logical connection to the feedback, and there is no alignment with their established learning profile. 2 - Poor The student's updating or use of their knowledge base is vague, contradictory, or lacks logical structure. Any observed changes do not clearly align with the student's profile. 3 - Generally Adequate The student shows some updating of their knowledge base and attempts to use it logically when responding to feedback. However, consistency with the student's profile is only partially observed—some steps fit. 4 - Accurate but Not Practical The student clearly updates and applies their knowledge base in a logically coherent way as they process feedback. The observed reasoning is broadly consistent with their profile. 5 - Excellent The student's knowledge base is observably updated and utilized in a clear, logical manner while processing feedback. Moreover, the updates and their application are fully consistent with their profile.

Table 15: Evaluation Questions and Scoring Criteria.

Participant Id	CPT-Agent					GLM-4.5		
	Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3
Participant 1	3.08	3.33	3.17	3.75	3.67	2.00	1.67	1.67
Participant 2	3.42	3.08	3.25	3.58	2.83	2.83	1.83	2.00
Participant 3	3.08	2.92	3.17	3.58	3.33	2.33	2.17	1.50
Participant 4	3.17	2.75	3.00	3.17	2.83	2.60	1.80	1.40
Participant 5	3.08	2.83	3.17	3.25	2.83	2.83	2.17	2.00
Average Score	3.17	2.98	3.15	3.47	3.10	2.52	1.93	1.71
Standard Deviation	0.14	0.23	0.09	0.25	0.38	0.36	0.23	0.28

Table 16: The average scores of simulated agents over 5 interview questions of the 5 participants.