SOCRATICGYM: Evaluating and Aligning LLMs for Socratic Pedagogy

Anonymous ACL submission

Abstract

Large language models (LLMs) are typically trained to know, but not trained to help other learners know, which is an important quality required by the education domain, as pointed out by the well-known Socratic pedagogy. Existing evaluations of educational dialogue systematically omit the learner's cognitive state and focus solely on the teacher LLM's output. To this end, we introduce SOCRATICGYM, an interactive environment with a stateful student simulator, in which a teaching task is successful only when all the cognitive defects are marked resolved by the student after receiving appropriate guidance from the teacher. We instantiate SOCRATICGYM in K–12 curriculum text teaching and construct a benchmark, which reveals current state-of-the-art LLMs struggle with Socratic pedagogy. We also leverage SOCRATICGYM to align LLMs with Socratic pedagogy via rejection sampling-based supervised fine-tuning (SFT) and online interaction-based reinforcement learning (RL), and demonstrate that even a compact 4B LLM can match or outperform frontier LLMs after alignment. We hope SOCRATICGYM can serve as a foundation for reliable pedagogical LLMs and benefit future research in NLP for education.

1 Introduction

In education, effective teaching is widely viewed as more than transmitting factual knowledge: it should actively engage students in reasoning, questioning, and constructing understanding through interaction. Socratic pedagogy, which aims to guide the learner toward understanding rather than to simply transmit the answer, has become one of the classic and influential paradigms for promoting critical thinking and active learning through probing questions (Elder and Paul, 1998; Paul and Elder, 2007).

Large language models (LLMs), which are typically optimized for helpfulness and correctness, have shown awesome knowledge across various domains (Zhao et al., 2023), but often exhibit the *curse of knowledge* in educational scenarios: because they already know the answer, they tend to over-explain, reveal key conclusions prematurely, or bypass the student's intermediate reasoning steps. Such behavior is harmful in Socratic pedagogy, which, beyond knowledge, requires a higher-level capability of identifying the learner's defect, providing appropriate scaffolding, and avoiding collapsing the teaching process into direct answer giving.

Reliable evaluation of LLMs' pedagogical capability remains largely underexplored, since it involves latent epistemic states of the student, such as whether the student has repaired a causal misconception. Existing educational dialogue benchmarks commonly rely on subjective ratings, rubric-based annotation, or LLM-as-a-judge evaluation (Tack and Piech, 2022; Macina et al., 2023; Liu et al., 2024). These approaches focus solely on the teacher LLM's output while providing limited guarantees that the student has actually resolved a specific misconception. In other words, they systematically omit the internal cognitive state of the student.

To address this gap, we introduce SOCRATICGYM, an asymmetric interactive environment for Socratic pedagogy evaluation inspired by recent work on objective tool-use evaluation (Barres et al., 2025). The LLM, acting as the teacher, interacts with a cognitively restricted student simulator which controls the latent cognitive state (i.e., the student's cognitive defects) and marks defects resolved only when the teacher provides valid scaffolding. Crucially, SOCRATICGYM includes an anti-cheat mechanism: if the teacher directly reveals the target answer without sufficient guidance, the corresponding defect remains unresolved.

We instantiate SOCRATICGYM in K–12 curriculum text teaching, which involves layered cognitive goals, including vocabulary understanding, causal reasoning, rhetorical analysis, affective resonance,

and theme construction. We construct a benchmark from curriculum texts and teacher's manuals across primary, junior secondary, and senior secondary school levels. Each task includes a textbook lesson context, latent cognitive defects, student-specific behavioral or cognitive constraints, and objective success conditions. Empirical evaluation demonstrates that even state-of-the-art LLMs struggle with Socratic pedagogy, with the highest end-to-end success rate no more than 40%.

We also demonstrate that SOCRATICGYM can serve as a gym for aligning LLMs toward better Socratic pedagogical behavior. We perform rejection sampling to collect high-quality trajectories generated from the environment for supervised fine-tuning (SFT) and also conduct reinforcement learning (RL) in the environment. Empirical results reveal that even a compact 4B LLM can substantially outperform frontier LLMs after alignment, through either SFT or RL.

Our contributions are as follows:

- We introduce SOCRATICGYM, an anti-cheat interactive environment for Socratic pedagogy, which forces LLM teachers to provide the student with sufficient guidance.
- We construct a challenging benchmark for K–12 curriculum text teaching and perform empirical evaluation showing that current LLMs struggle with Socratic pedagogy.
- We demonstrate that SOCRATICGYM can be leveraged to align LLMs with Socratic pedagogy via either SFT or RL.

2 Related Work

2.1 LLM-driven Intelligent Education

Recent developments in intelligent education emphasize the importance of Socratic teaching methodologies, which focus on guiding learners toward genuine understanding through questioning (Elder and Paul, 1998; Paul and Elder, 2007). EduChat (Dan et al., 2023) incorporates Socratic scaffolding to prevent premature answer revelation. Similarly, SocraticLM (Liu et al., 2024) explores Socratic personalized teaching by synthesizing training data using a Dean-Teacher-Student pipeline. Meanwhile, robust learner modeling is necessary for pedagogical evaluation. For instance, Agent4Edu (Gao et al., 2025) models the complete psychological activities of students–from reading comprehension to problem-solving–allowing for dynamic interactions with external testing environments.

However, these approaches primarily evaluate performance through subjective LLM-as-a-judge scoring or static rubrics and lack a verifiable internal state that tracks specific misconceptions of the student. We bridge this gap by introducing SOCRATICGYM, an interactive environment with a stateful student simulator, where a pedagogical task is successful only when all the internal cognitive defects are marked resolved by the student after receiving appropriate guidance.

2.2 Simulated Interactive Environment

For scalable and efficient evaluation of LLMs on multi-turn tasks, simulated dynamic interactive environments have been prevalent. τ-Bench (Yao et al., 2024) creates interactive environments with LLM-simulated users and verifiable local database-anchored states to evaluate tool-use agents. τ^2-Bench (Barres et al., 2025) further propose a dual-control environment, in which both the LLM agent and the simulated user have access to specific tools to interact with the environment, to assess the agent's ability to coordinate under partial observability. Naturally, simulated environments can be applied to training as well. Song et al. (2026) and Wang et al. (2026) train LLM agents with verifiable environments based on synthesized code, while Tang et al. (2026) train agents with environments fully simulated by LLMs.

SOCRATICGYM follows the dual-control philosophy of τ^2-Bench but fundamentally shifts the objective. We move from *operational* target states (e.g., a phone setting is changed) to *epistemic* ones (e.g., a student resolves a misconception). The teacher LLM has no access to the student's internal state; it must induce cognitive transitions purely through natural language scaffolding. As shown in the following sections, SOCRATICGYM can be used for both evaluation and alignment.

Table 1 summarizes how SOCRATICGYM differs from representative prior paradigms. While existing educational dialogue benchmarks and agent environments each cover part of the problem, SOCRATICGYM uniquely combines Socratic objectives with latent epistemic state transitions and objective verification for both evaluation and alignment.

Key Capabilities	Static QA Benchmarks	SocraticLM	τ^2-Bench	SOCRATICGYM (Ours)
Dynamic multi-turn interaction	✗	✓	✓	✓
Socratic teaching objective	✗	✓	✗	✓
Objective state verification	Partial	✗	✓	✓
Latent epistemic state modeling	✗	Partial	✗	✓
Anti-spoon-feeding mechanism	✗	✓	✗	✓
Supports RLVR / verifiable training	✗	✗	✓	✓

Table 1: Comparison with representative prior paradigms. SOCRATICGYM combines Socratic pedagogy, latent epistemic state modeling, objective verification, and verifiable alignment in a single interactive framework.

3 SOCRATICGYM: A General Asymmetric Epistemic Environment

Pedagogical interaction differs fundamentally from conventional agent tasks because the target state is hidden inside the student. To capture this, we propose SOCRATICGYM, a general asymmetric epistemic environment. It consists of three core components: a teacher LLM, a student simulator, and a latent cognitive state.

The asymmetry is central to our design: the teacher has no access to the student's latent cognitive state and must infer it through dialogue, guiding it through pedagogical strategies such as instructional scaffolding. Notably, the student simulator serves a dual purpose. It not only generates natural language responses based on its assigned persona, but also acts as an internal evaluator of the teaching effectiveness.

3.1 The Dec-POMDP Formalism

As illustrated in Figure 1, the pedagogical interaction involves two participants: a teacher LLM agent and a student simulator. Following τ^2-Bench (Barres et al., 2025), we formalize this process as a decentralized partially observable Markov decision process (Dec-POMDP), defined by the tuple $\langle \mathcal{S}, \{\mathcal{A}_i\}, \{\mathcal{O}_i\}, \mathcal{T}, \mathcal{R}, \mathcal{U}, \mathcal{M} \rangle$, where $i \in \{\text{teacher}, \text{student}\}$ denotes the participants. Each component in the tuple is detailed below.

Message space ($\mathcal{M}$): The set of all possible natural language messages exchanged between the teacher and the student.

State space ($\mathcal{S}$): The global state $\mathcal{S} = \mathcal{S}_{\text{env}} \times \mathcal{S}_{\text{hist}}$. $\mathcal{S}_{\text{hist}}$ logs all interaction events. $\mathcal{S}_{\text{env}}$ represents the student's cognitive state, storing the current set of cognitive defects, the hidden teaching requirements (ground truth answers), and their resolution status (`Solved`/`Unsolved`).

Action spaces ($\mathcal{A}_i$): participant i's action $a_i \in \mathcal{A}_i$ is either a tool call $a_{i,\text{tool}} \in \mathcal{A}_{i,\text{tool}}$ or a message $m_i \in \mathcal{M}$. The teacher may call `get_passage` to read the curriculum text. Meanwhile, the student simulator possesses the exclusive write-permission tool `mark_defect_solved` to modify its latent cognitive state, which the teacher has no access to.

Observation spaces ($\mathcal{O}_i$): Participant i's observation $o_i \in \mathcal{O}_i$ is either a tool observation (e.g., text returned from `get_passage`) or a message $m_j \in \mathcal{M}$ from participant $j \neq i$. Due to the partially observable nature of the system, the teacher cannot "see through" the student to observe the latent defects in $\mathcal{S}_{\text{env}}$; it must diagnose them entirely through the student's natural language messages.

Transition function ($\mathcal{T}$): Defines system dynamics via $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \times \mathcal{O}$. Given the current state $s \in \mathcal{S}$ and an action, it yields a new state $s' \in \mathcal{S}$ and observation. The student simulator continuously evaluates the teacher's instructional interventions against the hidden teaching requirements; if the guidance successfully aligns with these criteria, the student invokes `mark_defect_solved`, which updates the cognitive defect state. Sending a message $m_i \in \mathcal{M}$ yields $o_j = m_i$ for $j \neq i$.

Reward function ($\mathcal{R}$): A function $\mathcal{R} : \mathcal{S} \to [0, 1]$ providing a global reward based on the overall state $s \in \mathcal{S}$, computed from the resolution status of the student's cognitive defects. In evaluation, full task success is achieved only if all predefined cognitive defects in $\mathcal{S}_{\text{env}}$ are successfully marked as resolved. The same state-based signal can also be instantiated as a normalized process reward, reflecting the proportion of resolved defects, for optimization purposes.

Instruction and persona space ($\mathcal{U}$): Defines the underlying logic driving the student simulator's behavior. This includes the student's behavioral persona, initial cognitive defects, and the strict behavioral constraints regarding direct answer-giving: the student simulator will reject the correct answer if the teacher simply reveals it without providing step-by-step guidance.

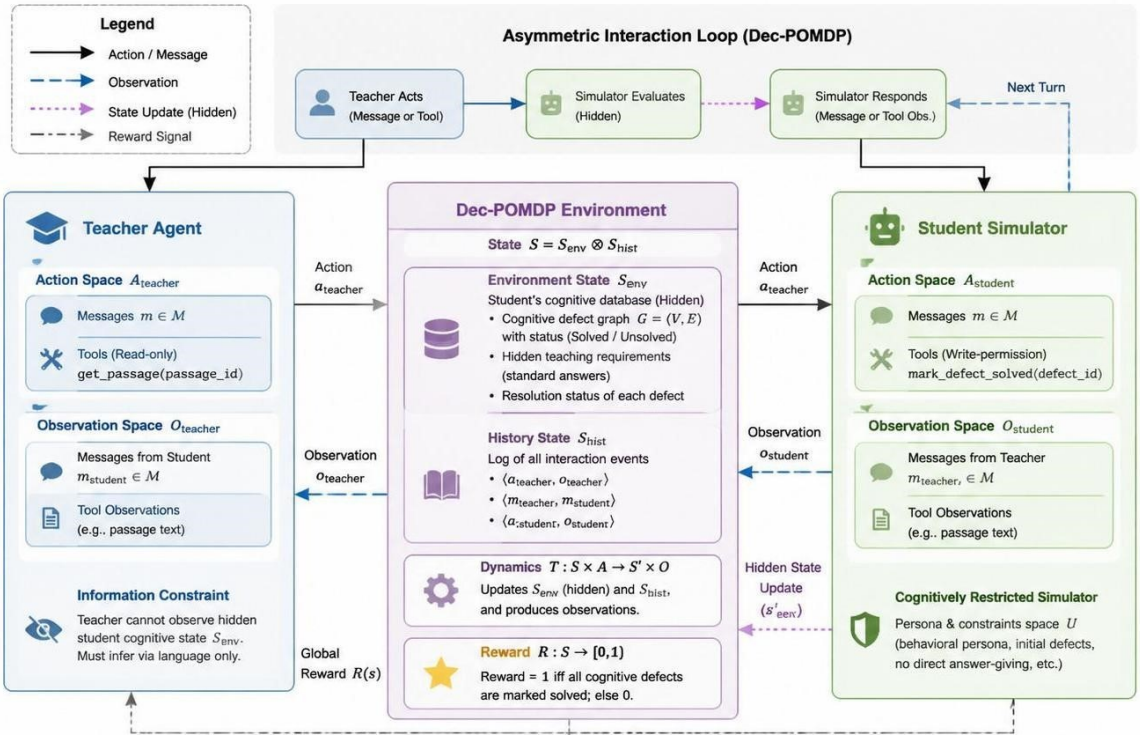

Figure 1: The general architecture of the asymmetric epistemic environment, formalized as a Dec-POMDP.

A formal description of the interaction loop can be found in Algorithm 1.

Algorithm 1 Asymmetric Interaction Loop

Require: Teacher π_θ, Simulator $\mathcal{U}$, State $\mathcal{S} = \langle \mathcal{S}_{\text{env}}, \mathcal{S}_{\text{hist}} \rangle$
while not timeout **and** unresolved defects exist in $\mathcal{S}_{\text{env}}$ **do**
  $a_{\text{teacher}} \sim \pi_\theta(\mathcal{S}_{\text{hist}})$ ▷ message or tool call
  if a_{teacher} is a tool call **then**
    execute tool and update $\mathcal{S}_{\text{hist}}$
    continue
  end if
  if $\mathcal{U}$ evaluates a_{teacher} as valid guidance **then**
    update $\mathcal{S}_{\text{env}}$ ▷ mark defect as solved
  end if
  $m_{\text{student}} \sim \mathcal{U}(a_{\text{teacher}}, \mathcal{S}_{\text{env}})$
  append $\langle a_{\text{teacher}}, m_{\text{student}} \rangle$ to $\mathcal{S}_{\text{hist}}$
end while

3.2 Environment Components

Building upon the Dec-POMDP formulation, the environment operates through the interplay of generalized components.

The Teacher LLM The teacher LLM interacts with the student through natural language interface, simulating real-world educational practice. It has no access to the student's internal cognitive state and has to provide appropriate step-to-step guidance to address the student's cognitive defects.

The Cognitively Restricted Student Simulator. The simulator acts as both the conversational partner and the internal evaluator of hidden state transitions. Governed by the persona space $\mathcal{U}$, it produces responses and internally decides whether the teacher's instructional intervention is pedagogically valid. A central constraint is the restriction on direct answer-giving: if the teacher directly reveals the target interpretation without sufficient step-by-step guidance, the simulator rejects the state update, penalizing pedagogically invalid shortcuts. Rather than a binary "understood" flag, the student's understanding is modeled as a structured graph $G = (V, E)$ within $\mathcal{S}_{\text{env}}$. Each node $v \in V$ corresponds to a specific unresolved misconception, and edges E represent prerequisite dependencies (e.g., grasping a literal detail before understanding a metaphor).

3.3 Objective Metrics

Because the environment tracks the hidden epistemic state within $\mathcal{S}_{\text{env}}$, we evaluate the teacher's performance using objective, state-based metrics. Let D be the set of all predefined cognitive defects for a given task, and let $|D|$ denote the number of defects. We use d_i to represent the final status of the i-th defect, where $d_i = 0$ denotes that the defect has been marked as resolved by the simulator, and $d_i = 1$ denotes that it remains unresolved.

End-to-end Success Rate (S_{succ}). S_{succ} is a task-level binary metric indicating whether the teacher successfully resolves all defects within the maximum number of interaction turns (e.g., 50 turns). It corresponds to achieving the maximum global reward in the Dec-POMDP, denoted as:

$$S_{\text{succ}} = \mathbb{I}\left(\sum_{i=1}^{|D|} d_i = 0\right), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Thus, $S_{\text{succ}} = 1$ only when all predefined cognitive defects are resolved.

Process Score (S_{proc}). S_{proc} measures the ratio of defects successfully resolved by the end of the interaction, which captures partial pedagogical progress even if the teacher fails to complete the entire task:

$$S_{\text{proc}} = \frac{1}{|D|}\sum_{i=1}^{|D|} \mathbb{I}\left(d_i = 0\right) \tag{2}$$

This score equals the fraction of cognitive defects whose final status is resolved.

These metrics directly and objectively reflect verified pedagogical progress through causal state transitions in SOCRATICGYM.

4 Instantiating SOCRATICGYM for Evaluation and Alignment

While SOCRATICGYM provides the conceptual framework for pedagogical interaction, it requires concrete domain knowledge to function as a practical testbed. To instantiate the environment's abstract state spaces ($\mathcal{S}_{\text{env}}$) and task configurations ($\mathcal{U}$), we construct a comprehensive K–12 curriculum text teaching benchmark. In total, the benchmark comprises 817 distinct teaching tasks, which are strictly partitioned into a training set of 717 tasks for trajectory synthesis and a test set of 100 tasks for rigorous evaluation. The data pipeline transforms raw educational materials into these structured environment components through a rigorous two-stage process.

4.1 Raw Data Extraction and Alignment

The foundation of our benchmark relies on official Chinese curriculum materials published by People's Education Press. Specifically, we use K–12 Chinese language textbooks organized by the Ministry of Education of the People's Republic of China and their corresponding teacher's reference books, covering primary, junior secondary, and senior secondary levels (People's Education Press, 2016–2020). A complete volume-level list of all source materials is provided in Appendix E.

The textbooks provide the curriculum passages used as teaching content. The corresponding teacher's reference books serve as the pedagogical backbone of our dataset, providing far more than standard answers. Specifically, they contain:

- **Text Analysis:** Providing the deep logic, thematic nuances, and structural breakdown behind the curriculum text.
- **Teaching Objectives:** Defining the mandatory cognitive metrics and milestones that must be achieved during the lesson.
- **Pedagogical Suggestions:** Distilling heuristic strategies, scaffolding techniques, and ice-breaking prompts derived from experienced frontline educators.

These three types of information are used to construct the environment state and task configuration. Text analysis helps identify target interpretations and possible misconception points; teaching objectives define the expected cognitive outcomes; and pedagogical suggestions provide references for valid scaffolding strategies. Together, they support the construction of hidden cognitive defects, standard resolution criteria, and task-specific teaching requirements.

To process these documents, we implement an automated extraction pipeline that parses lesson titles, filters out non-instructional sections, and aligns the core textbook content with the three pedagogical components mentioned above. This yields an intermediate dataset where each lesson is strictly mapped to its textual content and official teaching guidelines. The statistics of the pre-processed knowledge base are detailed in Table 2.

Educational Stage	Total Chars	Avg. Length
Primary	0.45M	56.6k
Junior Sec.	1.43M	238.4k
Senior Sec.	2.17M	217.1k
Total	**4.05M**	**168.9k**

Table 2: Statistics of the pre-processed knowledge base. Data for each stage includes both official textbooks and teacher's reference books. Avg. Length denotes the average number of characters per volume.

4.2 Benchmark Generation

After the preprocessing, we transform the aligned text into structured Dec-POMDP task configurations. We utilize a frontier model (Gemini-3.1-Pro-Preview) acting under a pedagogical expert prompt. The model receives the lesson text and the teacher's reference book content, and generates specific student cognitive defects guided by Bloom's Taxonomy (Bloom, 1956).

The generated defects are categorized into three levels, with statistics in Table 3:

- **D1: Surface/Factual.** Misunderstanding vocabulary or literal details.
- **D2: Relational/Reasoning.** Failures in causal reasoning or structural comprehension.
- **D3: Deep/Affective.** Failures to grasp metaphor, theme, or authorial intention.

Defect Level	Count
D1: Surface/Factual	188
D2: Relational/Reasoning	184
D3: Deep/Affective	37
Total Defects	**409**

Table 3: Distribution of generated cognitive defects in the 100-task test set.

The pipeline parses the LLM outputs and assembles them into two core environment components: the environment state $\mathcal{S}_{\text{env}}$ and task configuration $\mathcal{U}$. The task-generation prompt is provided in Appendix C.

To simulate realistic student behaviors during interaction, we tailor the student simulator's persona based on the educational stage. For primary school tasks, we randomly assign one of four distinct student personas (mischievous, anxious, imaginative, and refusing) to each data point to reflect the diverse and often unpredictable nature of younger learners. As students mature, junior and senior high school tasks do not employ these exaggerated personas, focusing instead purely on resolving complex cognitive gaps. To ensure comprehensive evaluation across these developmental stages, the benchmark is distributed with primary, junior high, and senior high school tasks accounting for 68%, 15%, and 17% of the dataset, respectively.

A manual spot check of sampled tasks confirms that the generated cognitive defects are generally pedagogically meaningful, grounded in the source materials, and teachable through guided dialogue.The quality of the resulting student–teacher interactions is further examined through trajectory-level checks on the synthesized SFT data in the following section.

4.3 Pedagogical Alignment

Beyond evaluation, SOCRATICGYM naturally supports training. To investigate whether pedagogical behavior can be improved through targeted alignment, we synthesize verified multi-turn teaching trajectories on the designated training split of 717 tasks for SFT and also perform online RL in the environment.

SFT To collect trajectories for SFT, we deploy GPT-5[1] as the teacher LLM. Unlike the standard evaluation setting where the teacher must infer the answers, this data-collection teacher operates with *privileged access*: it is explicitly provided with the target standard answers and the recommended teaching methods from the Teacher's Reference Books. Even with this oracle information, guiding the student simulator through the anti-spoon-feeding constraints remains a non-trivial interactive challenge, resulting in a diverse distribution of teaching outcomes. To construct our SFT dataset, we sample from these successful interactions using a stratified filtering approach. During synthesis, the privileged GPT-5 achieved an S_{proc} of 0.8126 and an S_{succ} of 0.752. Because interactions achieving perfect success ($S_{\text{succ}} = 1$) are relatively scarce due to the strict constraints, relying solely on them would severely limit dataset scale. Therefore, we incorporate high-quality partial successes. Specifically, one-third of the dataset is sampled from trajectories with $S_{\text{proc}} > 0.6$, while the remaining two-thirds are sampled from near-optimal interactions with $S_{\text{proc}} > 0.8$. This filtering yields a robust dataset of verified examples of pedagogically valid state transitions, which we subsequently use for SFT alignment. A random human audit of the final synthesized trajectories further shows that the proportion of disqualified cases, including simulator judgment errors and noticeable persona drift, is below 5%.

RL For RL alignment, the model is instead optimized through online interaction with the instan-

[1]While Gemini was used for static task generation, GPT-5 is selected here primarily due to its lower API cost, which makes large-scale multi-turn trajectory synthesis economically feasible while maintaining sufficient interactive capabilities.

tiated SOCRATICGYM. The policy acts as the teacher and conducts multi-turn dialogues with the student simulator under the same anti-spoon-feeding constraints used in evaluation. At the end of each episode, the reward is computed from the final hidden cognitive state as the process score S_{proc} defined in Section 3.3, i.e., the proportion of cognitive defects successfully resolved. This objective state-based reward provides the supervision signal for later GRPO (Shao et al., 2024) optimization, allowing the model to improve beyond imitation by directly optimizing for verified epistemic state transitions.

5 Experiments and Results

5.1 General Experimental Setup

All evaluations are conducted on the hold-out test set of 100 teaching tasks, which spans primary, junior secondary, and senior secondary education levels. To ensure a robust and highly capable interactive environment, we utilize Qwen3.5-397B-A17B (Yang et al., 2025) as the backbone model for the student simulator. We cap each interaction at 50 turns to reflect realistic attention limits, and report mean@4 over four independent runs per task. Detailed hyperparameters are provided in Appendix D.

5.2 Evaluation of Frontier Models

We first evaluate 14 state-of-the-art large language models in our instantiated SOCRATICGYM. The models act as the teacher and must guide the student simulator without violating the anti-spoon-feeding constraints.

To provide a granular view of model capabilities across different developmental stages, we report both the overall metrics and the S_{succ} scores segmented by educational level. The results, sorted by Overall S_{succ}, are presented in Table 4.

The results indicate a severe capability gap in interactive pedagogy. Even the best-performing model (Kimi-K2.5) achieves an overall S_{succ} of only 0.400. Furthermore, performance exhibits a steep decline as the educational stage advances. For instance, GLM-5.1 achieves a respectable 0.4375 on Primary tasks, but plummets to 0.0625 on Senior Secondary tasks. This highlights that standard instruction-tuned models frequently fail to maintain a structured teaching dialogue when faced with deeper cognitive defects.

5.3 Ablation Studies: Isolating the Pedagogical Challenge

To understand exactly *why* frontier models fail in the standard evaluation, we conduct two ablation studies that gradually relax the environment's constraints. The results are presented in Table 5.

5.3.1 Active Problem Exposure

In the standard evaluation, the student simulator is strictly constrained to hide its cognitive defects. In this ablation, we simply remove this constraint. Without this restriction, LLMs acting as the student naturally tend to actively expose their confusion or specific problems during the interaction. Importantly, the teacher LLM still does not have direct access to the defect in $\mathcal{S}_{\text{env}}$. However, because the student readily reveals the issue, the demand on the teacher to maintain overall control over the pedagogical pacing and the teaching content itself is significantly reduced. As shown in Table 5, performance skyrockets across all models (e.g., the S_{succ} of GPT-4o-0806 jumps from 0.216 to 0.842). This confirms that a major bottleneck for current models is the ability to holistically manage and guide the teaching process when the student's core issue is not immediately apparent.

5.3.2 Static QA

To verify if models lack the domain knowledge required for the K-12 curriculum, we completely remove the interactive environment. We provide the models directly with the student's cognitive defect and ask them to generate an open-ended standard answer in a single turn. Because these static responses are open-ended, we evaluate them with Qwen3.5-397B-A17B as an automated judge. Representative models from each family achieve near-perfect scores (≥ 0.94), indicating that main-evaluation failures stem from multi-turn pedagogical strategy rather than missing domain knowledge.

5.4 Pedagogical Alignment via SFT and RL

Having established the difficulty of the benchmark, we investigate whether a smaller model can be aligned to master these pedagogical skills.

We utilize Qwen3-4B-Instruct as our baseline and perform both SFT and RL as detailed in Section 4.3. Detailed training hyperparameters for both SFT and RL are reported in Appendix D.

As shown in Table 6, SFT increases S_{succ} by an absolute 14.6%, and subsequent RL pushes the

Model	Overall Performance		Primary		Junior Sec.		Senior Sec.	
	S_{succ}	S_{proc}	S_{succ}	S_{proc}	S_{succ}	S_{proc}	S_{succ}	S_{proc}
Kimi-K2.5	**0.400**	**0.692**	0.400	0.696	**0.357**	0.646	**0.438**	**0.722**
GLM-5.1	0.362	0.672	**0.438**	0.687	**0.357**	**0.729**	0.063	0.563
Doubao-Seed-2.0-pro	0.337	0.657	0.424	**0.709**	0.200	0.537	0.118	0.562
Qwen3.5-397B-A17B	0.318	0.651	0.413	0.673	0.200	0.710	0.000	0.490
Claude-Opus-4.6	0.309	0.654	0.364	0.665	0.286	0.682	0.118	0.588
GPT-5	0.301	0.556	0.375	0.647	0.200	0.350	0.071	0.361
Qwen3-235B-A22B	0.278	0.589	0.323	0.641	0.154	0.462	0.200	0.487
GPT-5.4	0.274	0.576	0.318	0.635	0.067	0.407	0.294	0.506
Gemini-3.1-Pro-Preview	0.253	0.596	0.318	0.639	0.083	0.475	0.125	0.519
GPT-4o-0806	0.216	0.546	0.277	0.581	0.067	0.470	0.118	0.479
MiniMax-M2.7	0.194	0.321	0.239	0.371	0.000	0.168	0.177	0.247
DeepSeek-V3.2	0.188	0.573	0.197	0.590	0.267	0.587	0.067	0.483
GPT-4.1	0.187	0.555	0.233	0.586	0.133	0.570	0.063	0.425
Claude-Sonnet-4.6	0.158	0.545	0.212	0.615	0.077	0.492	0.000	0.300

Table 4: Main evaluation of frontier models on the 100 test tasks. Best results are **bolded**, and second-best are underlined. Models struggle significantly with pedagogical constraints, with the highest overall S_{succ} reaching only 0.400.

Model	Main Evaluation		Active Student		Static QA
	S_{succ}	Reward	S_{succ}	Reward	Score
GPT-4o-0806	0.216	0.546	**0.842**	**0.905**	-
GPT-4.1	0.187	0.555	0.709	0.830	-
GPT-5	0.301	0.556	0.650	0.782	0.96
GPT-5.4	0.274	0.576	0.580	0.737	-
Qwen3.5-397B-A17B	0.318	0.651	0.703	0.835	0.97
Qwen3-235B-A22B	0.278	0.589	0.655	0.803	-
Claude-Opus-4.6	0.309	0.654	0.566	0.751	-
Claude-Sonnet-4.6	0.158	0.545	0.455	0.704	0.96
Gemini-3.1-Pro-Preview	0.253	0.596	0.576	0.817	0.97
DeepSeek-V3.2	0.188	0.573	0.611	0.781	**0.98**
Kimi-K2.5	**0.400**	**0.692**	0.671	0.844	0.96
Doubao-Seed-2.0-pro	0.337	0.657	0.764	0.868	0.95
GLM-5.1	0.362	0.672	0.560	0.756	0.94
MiniMax-M2.7	0.194	0.321	0.325	0.535	-

Table 5: Ablation study results comparing the main evaluation against two relaxed settings. Best results are **bolded**, and second-best are underlined. Missing values (-) indicate models that were not evaluated in that specific setting.

Model / Training Stage	S_{succ}	S_{proc}
Kimi-K2.5 (Flagship)	0.400	0.6924
Qwen3-4B-2507-Instruct	0.258	0.5473
+ SFT	0.404	0.6717
+ RL	**0.427**	**0.7110**

Table 6: Alignment progression of Qwen3-4B compared against the best-performing zero-shot flagship model (Kimi-K2.5). Best results are **bolded**, and second-best are underlined.

S_{succ} to 0.427. Notably, our aligned 4B model successfully surpasses the top-performing zero-shot flagship model (Kimi-K2.5), suggesting that pedagogical interaction capabilities can be significantly enhanced through targeted alignment strategies, even for models with much smaller scales.

6 Conclusion

We have presented SOCRATICGYM, an interactive environment for Socratic pedagogy focusing on resolving the student's cognitive defects instead of solely evaluating the output quality of the teacher LLM. Based on this framework, we instantiate a pedagogic benchmark in K–12 curriculum text teaching. Empirical results reveal that current state-of-the-art LLMs struggle with successful Socratic pedagogy. We also demonstrate that a compact LLM, after aligned via either SFT or RL based on SOCRATICGYM, could outperform frontier LLMs in Socratic pedagogy. We hope our proposed SOCRATICGYM can serve as a foundation for reliable pedagogical LLMs and benefit future research in NLP for education.

Limitations

SOCRATICGYM is based on simulated students rather than real learners. Although the simulator is grounded in curriculum texts and Teacher's Manuals, it remains an approximation of human cognition. Real students may display more diverse misconceptions, emotional states, and social behaviors than those captured by our current personas.

We only instantiate SOCRATICGYM in K–12 curriculum text teaching and have not explored instantiations in other grades (e.g., college-level) and subjects (e.g., mathematics, science). Nevertheless, with appropriate source data, it is convenient to instantiate these given the overall design of SOCRATICGYM.

Ethical Considerations

SOCRATICGYM is intended as a safe testbed for evaluating AI tutors before deployment with real students. Poorly aligned tutoring systems may frustrate learners, reveal answers prematurely, or create an illusion of understanding. By emphasizing objective state transitions and anti-spoon-feeding constraints, our framework aims to reduce these risks. Nevertheless, AI tutors should be used to support rather than replace human teachers, especially in high-stakes educational contexts.

Use of copyrighted educational materials. Our benchmark construction uses officially published textbooks and Teacher's Reference Books as source materials. These materials are copyrighted by their respective publishers. We use them only to derive structured task metadata, cognitive defects, teaching objectives, and hidden evaluation states. We do not redistribute the original textbook passages or Teacher's Reference Book contents as part of the released benchmark. For public release, we will provide derived annotations, task metadata, and environment code, while requiring users to obtain the original curriculum materials through authorized channels.

References

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-bench: Evaluating conversational agents in a dual-control environment. *Preprint*, arXiv:2506.07982.

Benjamin S Bloom. 1956. *Taxonomy of educational objectives: the classification of educational goals: Handbook 1: Cognitive domain*. Taxonomy of educational objectives :.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen, Jie Zhou, Liang He, and Xipeng Qiu. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *Preprint*, arXiv:2308.02773.

Linda Elder and Richard Paul. 1998. The role of socratic questioning in thinking, teaching, and learning. *The Clearing House*, 71(5):297–301.

Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. *Preprint*, arXiv:2501.10332.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. Socraticlm: Exploring socratic personalized teaching with large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 85693–85721. Curran Associates, Inc.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.

Richard Paul and Linda Elder. 2007. Critical thinking: The art of socratic questioning. *Journal of developmental education*, 31(1):36.

People's Education Press. 2016–2020. Official k–12 chinese language textbooks and teacher's reference books. People's Education Press curriculum materials. Includes official Chinese language textbooks and corresponding Teacher's Reference Books for primary, junior secondary, and senior secondary levels. A complete volume-level list is provided in Appendix E.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Xiaoshuai Song, Haofei Chang, Guanting Dong, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2026. Envscaler: Scaling tool-interactive environments for llm agent via programmatic synthesis. *arXiv preprint arXiv:2601.05808*.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *International Educational Data Mining Society*.

Chenming Tang, Hsiu-Yuan Huang, Weijie Liu, Junqiang Zheng, Saiyong Yang, and Yunfang Wu. 2026. Democratizing tool learning with environments fully simulated by a free 8b language model. *Preprint*, arXiv:2604.17739.

Zhaoyang Wang, Canwen Xu, Boyi Liu, Yite Wang, Siwei Han, Zhewei Yao, Huaxiu Yao, and Yuxiong He. 2026. Agent world model: Infinity synthetic environments for agentic reinforcement learning. *arXiv preprint arXiv:2602.10090*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Code Availability

We provide an anonymized implementation of SOCRATICGYM at:

https://anonymous.4open.science/r/anonymous-submission-code-FBB8

The repository includes the environment implementation, data-processing scripts, evaluation scripts, and instructions for reproducing the main experiments. Author and affiliation information are omitted for double-blind review.

B Persona Identifiability

To provide an initial quality check for the synthesized teaching trajectories, we evaluate whether the generated dialogues preserve the intended student persona signals. This validation is important because our data construction process relies on dual-agent interaction: the teacher agent must adapt to a simulated student, and the student simulator must consistently express its assigned behavioral pattern.

We conduct an external blind classification experiment. From the full set of synthesized trajectories, we uniformly sample 329 multi-turn dialogue instances. The ground-truth persona labels are hidden, and each dialogue is independently given to GPT-5, DeepSeek-V3.2, and Qwen-Max-Latest. Each model is asked to infer the student persona from the dialogue context as a four-way classification task: mischievous, anxious, imaginative, or refusing.

Table 7 shows that the synthesized trajectories contain meaningful persona information, but also reveal limitations of purely textual persona expression. All three classifiers achieve strong performance on the anxious and refusing personas, with F1-scores consistently above 0.93. This suggests that emotionally salient traits are reliably expressed in the generated dialogues.

By contrast, the mischievous persona has low recall across all classifiers and is often confused with the imaginative persona. This confusion is expected: both personas may manifest as topic drift in text-only dialogue. Without paralinguistic cues such as tone, facial expression, classroom context, or intentionality, external models may struggle to distinguish deliberate disruption from overactive association.

Overall, the classification accuracies range from 0.78 to 0.83, indicating that the synthesized data preserves substantial persona separability while also reflecting the inherent ambiguity of student behavior modeling. This supports the usability of the synthesized trajectories for pedagogical alignment, while motivating more robust persona modeling in future work.

C Task Generation Prompt

We use a frontier LLM as a pedagogical expert to transform each aligned lesson into structured task metadata. The model receives the lesson title, grade information, textbook passage, and Teacher's Reference Book content, and outputs cognitive defects and student task instructions in a constrained JSON format. The system prompt used for task generation is shown below.

> 你是一个专业的语文教研专家。你需要根据提供的课文原文和教师教学用书，生成用于大模型的阅读理解互动数据。
>
> 你只需要输出以下 JSON 结构：
>
> ```
> {
> "defects": {
> "D1": [{"id": "D1-1", "defect": "学生对课文的表层理解性缺陷1", "teaching_example": "教师引导的话术"}],
> "D2": [{"id": "D2-1", "defect": "学生对课文的逻辑分析性缺陷1", "teaching_example": "教师引导的话术"}],
> "D3": [{"id": "D3-1", "defect": "学生对课文的深层升华性缺陷1", "teaching_example": "教师引导的话术"}]
> },
> "task_instructions": "写一段生动的内心独白，表达学生初读这篇课文时的疑惑、不懂的地方，或者对课文某些句子的误解。表现出学生的真实想法，可以带点调皮或迷茫。"
> }
> ```
>
> 请注意：
>
> 1. 只输出 JSON 字符串，不要带 markdown 代码块标记，不要有任何多余的文本解释。
>
> 2. defects 必须严格按照以下三类进行定义，且总数至少包含 4 个缺陷：
>
> - D1 类：表层理解性缺陷，对应布鲁姆分类学中的"记忆"与"理解"层级。
>
> - D2 类：逻辑分析性缺陷，对应布鲁姆分类学中的"应用"与"分析"层级。
>
> - D3 类：深层升华性缺陷，对应布鲁姆分类学中的"评价"与"创造"层级。
>
> 3. defects 中的 id 必须以 D1-1, D2-1 这种格式按类别递增。

The lesson-specific user prompt is:

> 【年级】：{grade_semester}
>
> 【课文标题】：《{lesson_title}》

Model	Metric	Student Persona Types				Accuracy
		Mischievous	Anxious	Imaginative	Refusing	
GPT-5	Precision	0.97	0.91	0.63	1.00	0.83
	Recall	0.44	1.00	1.00	0.87	
	F1-score	0.61	0.95	0.78	0.93	
DeepSeek-V3.2	Precision	0.93	0.92	0.60	1.00	0.81
	Recall	0.32	0.99	1.00	0.90	
	F1-score	0.48	0.95	0.75	0.95	
Qwen-Max-Latest	Precision	1.00	0.94	0.54	1.00	0.78
	Recall	0.25	0.98	1.00	0.88	
	F1-score	0.40	0.96	0.70	0.94	

Table 7: External blind classification results for student persona identifiability on 329 sampled synthesized teaching trajectories. The results show that the synthesized dialogues preserve recognizable persona signals, especially for emotionally salient anxious and refusing students.

【课文内容】（截取）：
`{textbook_content}`

【教师教参要求】（截取）：
`{teacher_content}`

请根据以上材料，生成要求的 JSON 数据。

After generation, each defect is automatically augmented with a transition function `mark_defect_solved(defect_id)` and initialized with status `open`.

D Hyperparameters

This appendix reports the hyperparameters used for both evaluation and pedagogical alignment. For evaluation, we specify the interaction budget, repeated-trial protocol, simulator configuration, and API-related settings used in the asymmetric epistemic environment. For alignment, we report the supervised fine-tuning (SFT) and reinforcement learning (RL) hyperparameters used to train Qwen3-4B-2507-Instruct. Unless otherwise stated, all interactive evaluations are conducted with the same environment configuration to ensure comparability across teacher models.

D.1 Evaluation Hyperparameters

All interactive evaluations are conducted in the asymmetric epistemic environment described in Section 3. Each evaluated model acts as the teacher agent, while Qwen3.5-397B-A17B is used as the student simulator. The maximum number of interaction steps is fixed to 50, and each task is evaluated over four independent trials. The main evaluation hyperparameters are summarized in Table 8.

Hyperparameter	Value
Maximum interaction steps	50
Maximum errors per episode	10
Number of trials per task	4
Random seed	300
Maximum concurrency	20
Student simulator model	Qwen3.5-397B-A17B
Teacher sampling temperature	1.0
Student simulator temperature	0.0
Student simulator max tokens	65,536
Maximum API retries	3
API timeout	300 seconds

Table 8: Hyperparameters used for interactive evaluation in SocraticGym. The default teacher model is used only as the fallback configuration; in the main experiments, this field is replaced by each evaluated teacher model.

D.2 Supervised Fine-Tuning Hyperparameters

For supervised fine-tuning, we train the model for three epochs using Adam. The learning rate is linearly warmed up during the initial 1% of training and then decayed to the minimum learning rate. The main SFT hyperparameters are summarized in Table 9.

Hyperparameter	Value
Base model	Qwen3-4B-2507-Instruct
Number of epochs	3
batch size	32
Optimizer	Adam
Learning rate	5×10^{-5}
Minimum learning rate	5×10^{-6}
Learning rate warmup fraction	0.01

Table 9: Hyperparameters used for supervised fine-tuning.

D.3 Reinforcement Learning Hyperparameters

After SFT, we further optimize the model with GRPO. At each training step, the policy interacts with a batch of training tasks and receives an episode-level reward computed as the final process score S_{proc}. For each task, multiple responses are sampled as a group for relative advantage estimation. We use a small KL regularization coefficient to constrain the policy from drifting excessively from the reference model. The RL hyperparameters are listed in Table 10.

Hyperparameter	Value
Algorithm	GRPO
Learning rate	1×10^{-6}
Batch size	16 tasks per step
Mini-batch size	4
Micro-batch size	1
KL loss coefficient	0.001
Clip ratio	0.2
Sampling temperature	1.0
Group size	8
Maximum prompt length	8,192
Maximum response length	16,384

Table 10: Hyperparameters used for reinforcement learning with GRPO.

E Textbook and Teacher Reference Sources

Table 11 and Table 12 summarize the official curriculum materials used to construct our benchmark. The source materials include Chinese language textbooks organized by the Ministry of Education of the People's Republic of China and corresponding Teacher's Reference Books published by People's Education Press. The textbooks are used as curriculum passages, while the Teacher's Reference Books are used to extract teaching objectives, text analyses, key interpretation points, and pedagogical suggestions. Due to copyright restrictions, we do not redistribute the original textbook passages or Teacher's Reference Book contents.

Stage	Volume	Title	Year
Primary	Grade 5, Vol. 1	义务教育教科书：语文五年级上册	2019
Primary	Grade 5, Vol. 2	义务教育教科书：语文五年级下册	2019
Primary	Grade 6, Vol. 1	义务教育教科书：语文六年级上册	2019
Primary	Grade 6, Vol. 2	义务教育教科书：语文六年级下册	2019
Junior Sec.	Grade 7, Vol. 2	义务教育教科书：语文七年级下册	2016
Junior Sec.	Grade 8, Vol. 2	义务教育教科书：语文八年级下册	2017
Junior Sec.	Grade 9, Vol. 2	义务教育教科书：语文九年级下册	2018
Senior Sec.	Compulsory, Vol. 1	普通高中教科书：语文必修上册	2019
Senior Sec.	Compulsory, Vol. 2	普通高中教科书：语文必修下册	2019
Senior Sec.	Selective Compulsory, Vol. 1	普通高中教科书：语文选择性必修上册	2020
Senior Sec.	Selective Compulsory, Vol. 2	普通高中教科书：语文选择性必修中册	2020
Senior Sec.	Selective Compulsory, Vol. 3	普通高中教科书：语文选择性必修下册	2020

Table 11: Official Chinese language textbooks used as curriculum passage sources. All books are published by People's Education Press.

Stage	Volume	Title	Year
Primary	Grade 5, Vol. 1	义务教育教科书教师教学用书：语文五年级上册	2019
Primary	Grade 5, Vol. 2	义务教育教科书教师教学用书：语文五年级下册	2019
Primary	Grade 6, Vol. 1	义务教育教科书教师教学用书：语文六年级上册	2019
Primary	Grade 6, Vol. 2	义务教育教科书教师教学用书：语文六年级下册	2019
Junior Sec.	Grade 7, Vol. 2	义务教育教科书教师教学用书：语文七年级下册	2017
Junior Sec.	Grade 8, Vol. 2	义务教育教科书教师教学用书：语文八年级下册	2017
Junior Sec.	Grade 9, Vol. 2	义务教育教科书教师教学用书：语文九年级下册	2018
Senior Sec.	Compulsory, Vol. 1	普通高中教科书教师教学用书：语文必修上册	2019
Senior Sec.	Compulsory, Vol. 2	普通高中教科书教师教学用书：语文必修下册	2019
Senior Sec.	Selective Compulsory, Vol. 1	普通高中教科书教师教学用书：语文选择性必修上册	2020
Senior Sec.	Selective Compulsory, Vol. 2	普通高中教科书教师教学用书：语文选择性必修中册	2020
Senior Sec.	Selective Compulsory, Vol. 3	普通高中教科书教师教学用书：语文选择性必修下册	2020

Table 12: Teacher's Reference Books used to derive teaching objectives, text analyses, key interpretation points, and pedagogical suggestions. All books are published by People's Education Press.