

Beyond Human Preferences: Simulated-Student Preference Alignment for AI Tutors

Anonymous ACL submission

Abstract

Recent AI tutors are powered by large language models and their alignment are largely human-centric; relying on scarce, costly, and ethically constrained human preference data. Moreover, such alignment optimizes surface-level response quality and fails to reflect learner diversity. In this position paper, we identify challenges in current alignment methods and propose a framework that offers a scalable alternative to static human feedback by generating preference signals through interactions with diverse simulated students. This reframes alignment as a learner-conditioned optimization problem, enabling tutor policies to optimize for understanding, engagement, and productive struggle rather than surface-level response quality. The framework is compatible with modern preference-learning methods such as DPO, IPO, and KTO. Finally, a small scale study on algebra tutoring demonstrates that preference distributions vary systematically across learner profiles, highlighting the importance of learner-aware alignment while directly addressing data scarcity and scalability challenges in AI tutor alignment.

1 Introduction

Large language models (LLMs) are increasingly deployed as intelligent tutors, providing explanations, hints, and feedback across diverse learning contexts (Danry et al., 2023; Lee et al., 2025; Hou et al., 2025; Rogers et al., 2025; Park and Seo, 2025). Recent works align these systems with human preferences using methods such as reinforcement learning from human feedback and direct preference optimization to improve response quality, helpfulness, and safety (Scarlatos et al., 2025; Dinucu-Jianu et al., 2025; Sonkar et al., 2024b).

In practice, high-quality human preference data for tutoring is difficult to obtain (Sonkar

et al., 2024b; Wang et al., 2025). Even when available, such data can reflect annotator biases (Góral et al., 2025) and narrow pedagogical assumptions that fail to capture learner diversity (Chiner et al., 2015). Moreover, real student-tutor interaction data is expensive, ethically sensitive, and constrained by privacy requirements (Christiano et al., 2017; Kucharyav et al., 2025). As a result, alignment solely based on human judgments focus optimizing for surface-level qualities such as fluency or correctness rather than deeper learning outcomes including conceptual understanding, cognitive engagement, and productive struggle.

Meanwhile, advances in LLM-based agent simulation have enabled generative student agents that exhibit plausible learning behaviors, reasoning patterns, and misconceptions (Gao et al., 2025; Xu et al., 2024). Prior works have demonstrated the potential of such simulated learners for data generation (Gao et al., 2025), educational system testing (Weitekamp et al., 2025), and classroom behavior modeling (Xu et al., 2025). However, existing approaches treat simulated students primarily as passive data sources or evaluation tools, rather than as active participants in alignment.

In this position paper, we argue that simulated students (Park et al., 2023) can serve as scalable and pedagogically meaningful preference signal generators for aligning AI tutors. We propose *Simulated-Student Preference Alignment* (SSPA), in which preferences are derived from interactions with diverse, parameterized student agents. SSPA reframes alignment as a learner-conditioned optimization problem, shifting the objective from producing responses that sound good to responses that support learning for different types of students. Rather than replacing human oversight, SSPA offers a complementary approach that en-

085	ables scalable exploration of pedagogical strate-	and preference datasets with minimal human	134
086	gies, robust evaluation under diverse learner	supervision (Seo et al., 2025; Scarlatos et al.,	135
087	conditions, and reduced reliance on sensitive	2024; Sonkar et al., 2024b). Recent work fur-	136
088	real-student data.	ther leverages simulated students to evaluate	137
089	This paper makes three contributions: (1)	tutor responses and guide optimization objec-	138
090	we identify a fundamental mismatch between	tives. For example, prior work trains tutors by	139
091	human-centric alignment methods and the	scoring candidate utterances using predicted	140
092	goals of educational AI; (2) we formalize SSPA	student correctness from a student model com-	141
093	as a general framework compatible with mod-	combined with rubric-based pedagogical evalua-	142
094	ern preference-learning methods including Di-	(Scarlatos et al., 2025). These approaches move	143
095	rect Preference Optimization (DPO) (Rafailov	toward outcome-aware optimization while pre-	144
096	et al., 2023), Identity Preference Optimization	servicing scalability.	145
097	(IPO) (Azar et al., 2024), and Kahneman Tver-		
098	sky Optimization (KTO) (Ethayarajh et al.,	2.3 Modeling Student Cognition,	146
099	2024); and (3) we outline key design choices,	Motivation, and Learning	147
100	benefits, limitations, and open challenges con-	Dynamics	148
101	cerning validity, bias, and evaluation. We hope	Recent work has significantly improved the fi-	149
102	this perspective encourages the community to	delity of simulated students by modeling key	150
103	rethink how learning-oriented alignment should	aspects of human learning. LLM-based simu-	151
104	be defined, operationalized, and evaluated in	lators are now conditioned on cognitive abil-	152
105	the era of generative AI tutors.	ity, motivation, engagement, personality traits,	153
		and problem-solving styles, enabling diverse	154
106	2 Background	and human-like learner behaviors (Liu et al.,	155
		2024; Xu et al., 2024; Gao et al., 2025; Pan	156
107	2.1 Learning from Human Preferences	et al., 2025). Simulators have also been im-	157
108	for Optimizing AI Tutors	proved to capture long-term learning dynam-	158
		ics such as retention, forgetting, and sustained	159
109	Learning from Human Preferences (LHP) is	versus brittle knowledge acquisition (Yuan et al.,	160
110	a dominant paradigm for aligning LLMs with	2025; Gao et al., 2025; Wu et al., 2025), demon-	161
111	desired behaviors, using techniques such as	strating that short-term correctness is a weak	162
112	reinforcement learning from human feedback,	proxy for learning.	163
113	reward modeling, and direct preference opti-	Several studies further model systematic	164
114	mization (Christiano et al., 2017; Ouyang et al.,	student misconceptions through reasoning-	165
115	2022; Rafailov et al., 2023; Bai et al., 2022).	consistent incorrect solutions and structured	166
116	These methods are widely used to improve the	error patterns (Sonkar et al., 2024a; Ross and	167
117	helpfulness, safety, and usability of conversa-	Andreas, 2025), enabling systematic evalua-	168
118	tional agents.	tion of tutor behavior and instructional strate-	169
119	In educational settings, preference-based	gies. These advances motivate our proposal to	170
120	learning optimizes AI tutors using judgments	treat simulated learners as active preference	171
121	from expert educators, instructional designers,	generators for aligning AI tutors.	172
122	or teaching assistants who rank or select pre-		
123	ferred responses (Woodrow et al., 2025; Scar-	3 Position: Simulated Students as	173
124	latos et al., 2024). These approaches encode	Preference Signal Generators	174
125	pedagogical norms such as correctness, clarity,	Current preference-based alignment methods	175
126	and politeness and implicitly assume that ex-	for AI tutors have proven effective for improv-	176
127	pert judgments correspond to improved student	ing surface-level instructional qualities such as	177
128	learning.	correctness and clarity. In this section, we	178
129	2.2 Synthetic Data for AI Tutor	argue that the standard paradigm of align-	179
130	Optimization	ing LLMs with static human preferences is	180
		fundamentally insufficient for the develop-	181
131	Advances in LLMs have accelerated synthetic	ment of intelligent tutoring systems and	182
132	data generation for education, enabling auto-	identify three core constraints that necessitate	183
133	mated feedback, synthetic tutoring dialogues,	a shift toward	

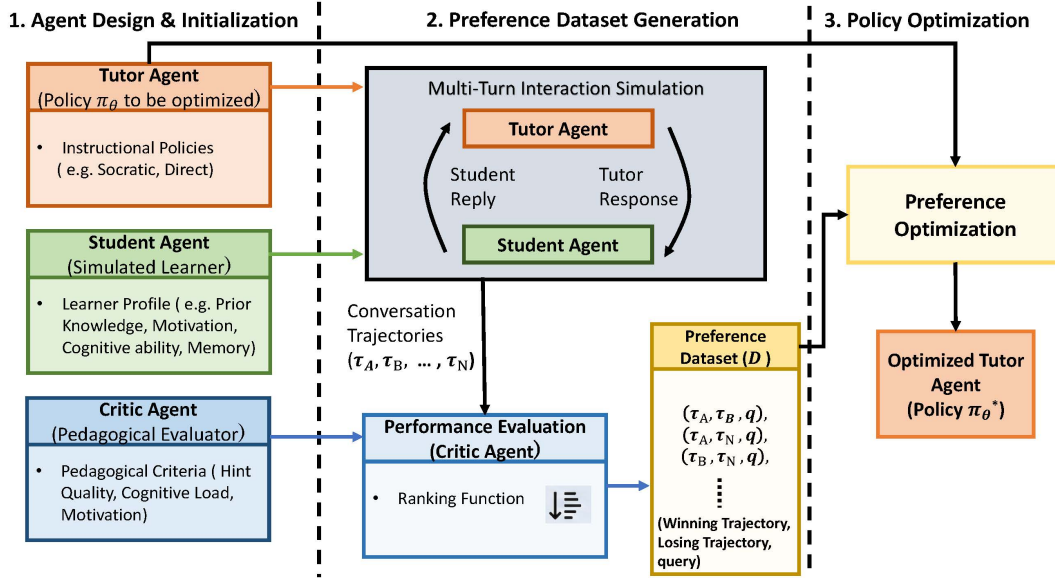


Figure 1: Overview of the proposed framework. The framework comprises three stages: (1) agent design and initialization, including a tutor agent, a simulated student agent with a learner profile, and a pedagogical critic; (2) multi-turn tutor–student interaction simulation, where conversation trajectories are evaluated and ranked to construct a preference dataset; and (3) tutor policy optimization based on the generated preferences.

184 learner-centric alignment using simulated stu- 211
 185 dents. 212

186 3.1 Constraints of Human-Centered 213 187 Preference Collection 214

188 Human preference optimization is widely used 215
 189 to improve language model outputs of AI tu- 216
 190 tors, typically relying on instructors, teach- 217
 191 ing assistants, or expert annotators to evalu- 218
 192 ate isolated tutor responses. While valuable, 219
 193 this approach is costly, sparse, and ethically 220
 194 constrained (Piech et al., 2015; Sonkar et al., 221
 195 2024b). Another issue is that, the quality of 222
 196 human-annotated preference datasets can vary, 223
 197 as models trained on human feedback may in- 224
 198 herit annotator biases or inconsistencies. For 225
 199 example, analyses of widely used preference 226
 200 datasets have shown that a substantial frac- 227
 201 tion of preferred responses contain factual or 228
 202 reasoning errors (Góral et al., 2025). 229

203 Moreover, collecting real student–tutor inter- 230
 204 action data at scale requires institutional ap- 231
 205 proval, informed consent, and careful handling 232
 206 of sensitive information. Even when available, 233
 207 such data is often sparse and difficult to anno- 234
 208 tate with reliable preference labels that reflect 235
 209 learning outcomes rather than superficial quali- 236
 210 ties. As a result, existing systems typically rely

211 on relatively small preference datasets collected 212
 213 from limited numbers of instructors or teaching 214
 215 assistants (Woodrow et al., 2025; Wang et al., 216
 217 2025). 218

219 3.2 Learning Outcomes Beyond 220 221 Correctness 222

223 To address scalability, recent work explores syn- 224
 225 thetic preference generation using automated 226
 227 critics, rubric-based scoring, or student models 228
 229 (Scarlatos et al., 2024; Seo et al., 2025; Sonkar 230
 231 et al., 2024a). While this improves scale, pref- 232
 233 erence signals often emphasize correctness and 234
 235 pedagogical form rather than learning dynam- 236
 237 ics. Even when simulated students are involved, 238
 239 they are typically passive evaluators or data 239
 240 generators (Scarlatos et al., 2024), leaving opti- 241
 242 mization detached from how learners respond, 242
 243 struggle, or improve over time. In other words, 243
 244 while synthetic data improves scalability, it 244
 245 does not fundamentally change what is being 245
 246 optimized. 246

247 3.3 Tutor-Centric vs. Learner-Centric 248 249 Alignment Objectives 249

250 Formally, given a student query q and candi- 251
 252 date responses $\{r_1, \dots, r_n\}$, preference learning 252
 253 optimizes a human-defined function $P_{\text{human}}(r | 254
 255 q)$. 255
 256 256

237 *q*). This assumes that human-preferred re- 286
 238 sponses promote learning, while educational 287
 239 research contradicts this assumption. Edu- 288
 240 cational feedback preferences are inherently 289
 241 heterogeneous. Effective feedback depends on 290
 242 learner characteristics such as prior knowledge, 291
 243 misconceptions, motivation, and cognitive load 292
 244 (Piech et al., 2015; Holstein et al., 2018). A 293
 245 response optimal for one learner may hinder an- 294
 246 other. Thus, a single global preference function 295
 247 fails to capture learning heterogeneity, moti- 296
 248 vating learner-conditioned alignment. 297

249 3.4 Problem Statement 298

250 Taken together, these observations highlight 300
 251 a central limitation of current approaches: 301
 252 human-centric alignment objectives are not 302
 253 enough for optimizing AI tutors for learning. 303
 254 Effective tutoring requires optimization with re-
 255 spect to learner-conditioned outcomes shaped
 256 by student state and interaction dynamics. 304
 257 This motivates the need for alternative align-
 258 ment paradigms that (1) account for learner 305
 259 diversity, (2) provide scalable and ethically vi- 306
 260 able preference signals, and (3) explicitly target 307
 261 learning-oriented objectives. In the following 308
 262 section, we introduce Simulated-Student Prefer-
 263 ence Alignment (SSPA) as one such paradigm, 309
 264 reframing alignment as a learner-centric op-
 265 timization problem grounded in interactions
 266 with generative student agents. 310

267 4 Proposed Framework: 311

268 Simulated-Student Preference 312

269 Alignment 313

270 4.1 Paradigm Overview 314

271 We propose SSPA, a learner-centric alignment 315
 272 paradigm in which AI tutors are optimized 316
 273 using preference signals derived from interac- 317
 274 tions with simulated students rather than ex- 318
 275 clusively from human annotators. The core 319
 276 idea is to evaluate tutor responses based on the 320
 277 impact on learning-related outcomes, rather 321
 278 than surface-level response quality. 322

279 In contrast to conventional alignment 323
 280 pipelines, where tutor responses are ranked ac- 324
 281 cording to static human-defined criteria, SSPA 325
 282 reframes alignment as a *learner-conditioned* 326
 283 *optimization problem*. Tutor behavior is opti- 327
 284 mized with respect to simulated students, repre- 328
 285 senting diverse learning profiles. These profiles 329

286 encode variation in prior knowledge, cognitive 287
 288 style, motivation, and misconceptions. These 288
 289 allow the alignment objective to reflect hetero- 289
 290 geneous learning needs. 290

291 Importantly, SSPA does not require modi- 291
 292 fying existing preference-learning algorithms. 292
 293 Instead, it changes the source and semantics 293
 294 of preference signals. Preferences are gener- 294
 295 ated through simulated learning interactions 295
 296 and evaluated according to pedagogical crite- 296
 297 ria. As such, SSPA is compatible with mod- 297
 298 ern preference-optimization methods while in- 298
 299 troducing a fundamentally different alignment 299
 300 target. Figure 1 provides an overview of the 300
 301 SSPA pipeline, illustrating how simulated tu- 301
 302 tor–student interactions are used to generate 302
 303 preference datasets that support tutor policy 303
 304 optimization. 304

304 4.2 Agent Roles and Interaction Loop 304

305 SSPA is instantiated using three agents: a **Stu-** 305
 306 **dent Agent**, a **Tutor Agent**, and a **Prefer-** 306
 307 **ence Evaluator**. Together, these agents 307
 308 define a closed-loop alignment process as de- 308
 309 scribed in figure 2. 309

310 **Student Agent.** The student agent mod- 310
 311 els a learner with a parameterized profile that 311
 312 captures long-term traits (e.g., prior knowl- 312
 313 edge, need for cognition, motivation, memory) 313
 314 and short-term state (e.g., confusion or en- 314
 315 gagement). Given a learning task or question, 315
 316 the student agent processes candidate tutor re- 316
 317 sponses, updates its internal state, and exhibits 317
 318 learning-relevant behaviors such as partial un- 318
 319 derstanding, misconceptions, or follow-up ques- 319
 320 tions. 320

321 **Tutor Agent.** The tutor agent generates 321
 322 multiple candidate responses for a given stu- 322
 323 dent query. These responses may differ in in- 323
 324 structional strategy, such as direct explanation, 324
 325 hint-based guidance, Socratic questioning, or 325
 326 motivational scaffolding. The goal of the tutor 326
 327 agent is not merely to produce a correct an- 327
 328 swer, but to explore pedagogical alternatives 328
 329 whose effectiveness depends on the learner. 329

330 **Preference Evaluator.** The preference eval- 330
 331 uator assesses the impact of tutor responses 331
 332 on the simulated student. Rather than judg- 332
 333 ing responses in isolation, it evaluates them 333
 334 in the context of student interaction, estimat- 334

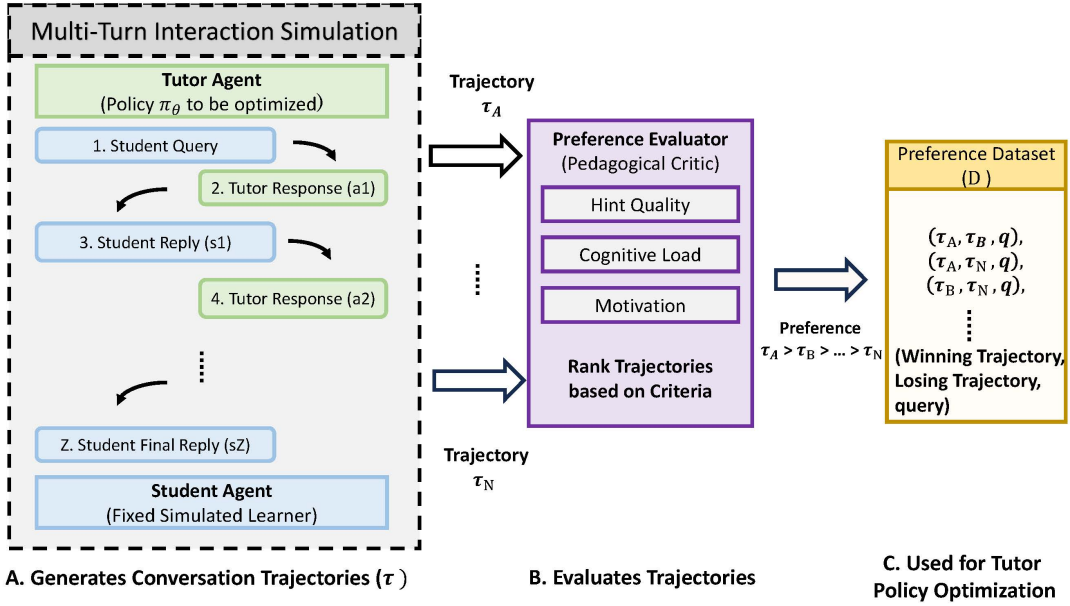


Figure 2: Overview of the preference dataset generation process. (A) generates multi-turn conversation between tutor and student; (B) critic agent evaluates the multi-turn conversation and ranks based on specified criteria; (C) based on the ranking given by critic agent generate pairwise preference dataset.

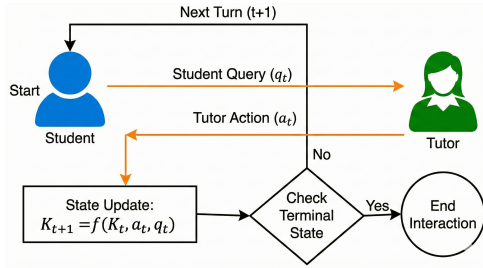


Figure 3: Overview of Multi-turn conversation between student and tutor. Student may initiate the conversation and based on student’s query the tutor gives feedback based on student’s profile. During the conversation student’s learning state updates.

ing which responses better support learning-oriented objectives. This evaluation yields pairwise or ranked preferences over tutor responses, conditioned on the student profile and interaction trajectory.

The interaction loop proceeds as follows: a simulated student initiates a learning dialogue by posing a query or response based on its current internal state. Given this student turn, the tutor generates candidate instructional response. The student agent then processes candidate response within a multi-turn dialogue, updating its internal learning state (e.g., understanding, confusion, or engagement) (see fig. 3).

The preference evaluator then compares the

resulting student state trajectories across alternative tutor responses and produces learner-conditioned preference judgments. The evaluator may be implemented as a separate model or as an explicit scoring function.

4.3 Preference Signal Design

A key design choice in SSPA is the definition of preference criteria. Unlike human preference learning, which often emphasizes fluency or correctness, SSPA focuses on *learning-oriented signals* that reflect pedagogical effectiveness. We highlight three illustrative criteria below.

Hint Quality. Preferences may favor responses that provide appropriate levels of guidance, neither revealing the full solution prematurely nor offering vague feedback that fails to advance understanding. Effective hints should support incremental reasoning and encourage active problem-solving.

Cognitive Load. Responses can be evaluated based on the cognitive demands they place on the learner. Overly complex explanations may overwhelm students with limited prior knowledge, while oversimplified responses may fail to challenge advanced learners. Preferences therefore reflect the alignment between response complexity and the student’s cognitive capacity.

Motivational Support. Learning is influenced not only by content but also by affective factors such as encouragement, confidence-building, and responsiveness to struggle. Preferences may favor responses that sustain engagement and reduce frustration, particularly for students exhibiting low motivation or repeated errors.

These criteria are not exhaustive, nor are they assumed to be universally optimal. Rather, they illustrate how preference signals in SSPA can be explicitly tied to pedagogical goals and conditioned on learner characteristics. By encoding such signals into the alignment objective, SSPA enables tutors to adapt instructional strategies to diverse learners instead of optimizing for a single, global notion of response quality. These signals can be used as pairwise comparisons over tutor responses. Figure 2 illustrates the preference dataset generation process in SSPA, showing how multi-turn tutor–student interactions are evaluated and transformed into pairwise preference signals for tutor optimization.

5 Proof-of-Concept Demonstration

To examine the practical feasibility of SSPA, we conducted a small-scale experiment that instantiates the alignment pipeline described in Section 4. The objective of this experiment is not to provide comprehensive empirical validation, but to demonstrate that the proposed framework can be implemented end-to-end and it produces meaningful learner-conditioned preference signals.

5.1 Experimental Setup

We conducted a small-scale study using ten algebra problems, three simulated learner profiles (novice, intermediate, advanced), and three tutoring strategies (direct explanation, hint-based guidance, Socratic questioning). Each configuration produced a short multi-turn tutor–student interaction.

Each interaction was evaluated by a preference evaluator that scores three learning-oriented signals: final answer correctness, estimated student understanding, and estimated student confusion. These signals operationalize the pedagogical criteria described in Section 4.3. The resulting scores were converted into scalar

Student Type	Direct	Hint	Socratic
Novice	18	8	4
Intermediate	12	10	8
Advanced	13	10	7

Table 1: Number of pairwise wins for each instructional strategy across three learner profiles and teaching strategies. Results indicate that preference distributions vary by learner type, supporting learner-conditioned alignment.

utilities to rank strategies and construct pairwise preference data for each learner profile as described in Appendix A.5. Full prompt specifications and example outputs are provided in Appendix A.

5.2 Results and Analysis

Table 1 reports the number of times each instructional strategy outperforms the others in pairwise comparisons across learner profiles. Although the direct strategy achieves the most wins overall, the relative preference distributions vary systematically with learner type. This indicates that preference signals are conditioned on learner characteristics rather than reflecting a single global notion of response quality. Despite the limited scale of the experiment, these patterns support our central claim that effective alignment for AI tutors must account for learner heterogeneity and interaction dynamics. The results provide initial evidence that SSPA produces coherent, learner-conditioned alignment signals and that the proposed framework is practically feasible.

While the study confirms the feasibility of SSPA, it also exposes some limitations. The generated preference signals are sensitive to the design of simulated learner personas, interaction length, and critic judgments. For example, novice learners sometimes fail to converge to correct solutions within short dialogue horizons, and the critic’s estimates of understanding and confusion may reflect linguistic confidence rather than deep conceptual mastery (see Appendix A.7.2). These behaviors highlight the risks of simulator dependence and noisy learning proxies, motivating the validity safeguards, uncertainty-aware filtering, confidence thresholds, and adversarial testing introduced in the next section.

Algorithm 1 Robust Preference Dataset Construction for SSPA (Validity & Safety Checks)

Require: Tutor policy π_θ ; simulated student population \mathcal{S} ; evaluator E ; task set \mathcal{Q} ; rollouts K ; thresholds $\tau_{\text{conf}}, \tau_{\text{agree}}$; adversarial tests \mathcal{A}
Ensure: Filtered preference dataset \mathcal{D}

```
1:  $\mathcal{D} \leftarrow \emptyset$ 
2: for each task  $q \in \mathcal{Q}$  do
3:   for each student  $s \in \mathcal{S}$  do
4:     Sample trajectories  $\{\tau_1, \dots, \tau_K\}$  by interacting  $\pi_\theta$  with  $s$ 
5:     Obtain utilities and uncertainty  $\{(u_i, c_i)\}_{i=1}^K \leftarrow E(\{\tau_i\}, s, q)$ 
6:     Select  $\tau^+, \tau^-$  by ranking  $u_i$ 
7:      $\Delta \leftarrow u(\tau^+) - u(\tau^-)$ ;  $c \leftarrow \min(c^+, c^-)$ 
8:     if  $c < \tau_{\text{conf}}$  or  $\Delta \leq 0$  then
9:       continue
10:    end if
11:    for each adversarial test  $a_j \in \mathcal{A}$  do
12:      Construct rephrased prompt  $p^{(j)} \leftarrow a_j(p)$ 
13:      if  $E(\tau^+, \tau^- | p^{(j)}, q) \neq \tau^+$  then
14:        continue 2
15:      end if
16:    end for
17:     $\mathcal{D} \leftarrow \mathcal{D} \cup \{(q, s, \tau^+, \tau^-)\}$ 
18:  end for
19: end for
return  $\mathcal{D}$ 
```

6 Evaluation Challenges and Validity Considerations

Using simulated students as preference generators raises important questions about validity and reliability. A primary concern is that preference signals may reflect the idiosyncrasies or limitations of the student model rather than generalizable learning principles. Prior work has noted that language-model-based simulators can produce brittle or overly consistent behaviors, raising concerns about overfitting tutor policies to synthetic learners (Yuan et al., 2025). To mitigate these risks, we emphasize the importance of uncertainty-aware filtering (Banerjee et al., 2025; Lu et al., 2025), where preference pairs with low simulator confidence or high disagreement across personas are discarded.

Another challenge is reward hacking (Taylor et al., 2025), where tutors learn to exploit simulator artifacts rather than support learning. Adversarial testing, such as introducing counterfactual student behaviors or alternative simulators, can help detect such failures.

These concerns are addressed concretely in Algorithm 1, which enforces uncertainty-aware filtering by removing low-confidence or weak

preference pairs and applies adversarial testing to detect reward hacking and simulator overfitting through counterfactual prompt variations that introduce small perturbations of the same learner conditions (e.g., rephrasing). Only preference pairs that satisfy these reliability and robustness constraints are retained, ensuring that the resulting dataset reflects stable, learning-oriented signals. Finally, transparency in simulator design and explicit documentation of modeled cognitive and motivational assumptions are essential for ethical deployment and reproducibility.

7 Benefits and Use Cases

SSPA offers several advantages over human-centric alignment approaches in educational settings, where learning outcomes are heterogeneous and latent. By positioning simulated learners as active participants in alignment, SSPA enables forms of tutor optimization and evaluation that are difficult to achieve using human preference data alone.

Reduced Reliance on Sensitive Real-Student Data. Educational data is constrained by privacy, consent, and institutional requirements (Piech et al., 2015). SSPA mitigates these limitations by shifting substantial portion of the alignment process to simulated environments, reducing dependence on sensitive real-student interactions.

Scalability and Coverage of Learner Diversity. Unlike human preference collection, which is costly and limited by ethical and logistical constraints (Christiano et al., 2017; Ouyang et al., 2022), simulated students can be instantiated at scale with controlled variation in learning profiles (Park et al., 2023; Gao et al., 2025; Xu et al., 2024). This enables systematic exploration of pedagogical strategies across differences in prior knowledge, cognitive capacity, motivation, and misconceptions, and supports stress-testing tutors under edge cases such as persistent misunderstanding or low engagement. By exposing tutors to these scenarios during alignment, SSPA encourages robustness and adaptability rather than optimization for an “average” learner.

Learner-Conditioned Pedagogical Adaptation. Since preference signals are generated in the context of diverse student

544 profiles, SSPA enables tutors to differentiate
545 instructional strategies based on learner needs.
546 The same problem may elicit different preferred
547 responses for novice versus advanced learners,
548 or for students with differing motivation levels
549 (Chi et al., 1989; Alevan et al., 2003). This
550 allows tutors to internalize pedagogical trade-
551 offs and adapt behavior dynamically, aligning
552 with principles of personalized instruction.

553 **Support for Curriculum-Level and Longitudinal Analysis.** As simulated students
554 can maintain internal state and memory across
555 interactions (Gao et al., 2025; Wang et al.,
556 2023), SSPA naturally extends beyond single-
557 turn tutoring. It supports alignment objectives
558 that account for learning trajectories over time,
559 including cumulative understanding, retention,
560 and curriculum progression. This opens op-
561 portunities for optimizing tutors not only for
562 immediate responses, but also for long term
563 educational goals.
564

565 8 Risks and Open Challenges

566 While simulated-student preference alignment
567 offers promising advantages, it also introduces
568 important limitations and risks that must be
569 carefully considered. Addressing these chal-
570 lenges is essential for ensuring that learner-
571 centric alignment improves educational out-
572 comes rather than introducing new sources of
573 bias or misalignment.

574 **Validity and Bias in Simulated Student Models.** A central concern is whether simu-
575 lated students faithfully represent real learners.
576 LLM-based student agents inevitably inherit
577 assumptions, limitations, and implicit biases
578 from their underlying models (Sheng et al.,
579 2019). If simulated learners fail to capture
580 key aspects of human learning such as affec-
581 tive states, social context, or metacognitive
582 processes, then the resulting preference signals
583 may misrepresent true pedagogical effectiveness
584 and favor dominant cultural or instructional
585 norms. This creates a risk of *simulation over-*
586 *fitting*, where tutors become optimized for arti-
587 facts of the simulator rather than real student
588 needs. Mitigating this risk requires validation
589 against human data, use of diverse simulators,
590 and systematic auditing of learner populations.

591 **Challenges in Defining Learning-Oriented Preference Signals.** Unlike
592
593

594 surface-level response quality, learning out-
595 comes such as understanding and engagement
596 are inherently latent and difficult to opera-
597 tionalize (Verma et al., 2023). Preference cri-
598 teria such as cognitive load or motivational
599 support are approximations, and their measure-
600 ment within simulated environments may be
601 noisy and indirect. Poorly specified preference
602 criteria may introduce misalignment by opti-
603 mizing for signals that weakly correlate with
604 real learning. Developing principled, theory-
605 informed learning metrics remains an open re-
606 search challenge.

607 **Role of Human Oversight and Evaluation.** SSPA is not a replacement for human
608 judgment. Human educators and learners re-
609 main essential for evaluating real-world effec-
610 tiveness, identifying failure modes, and guid-
611 ing ethical deployment (Bai et al., 2022; Park
612 et al., 2023). A key challenge is integrating
613 simulated-student alignment with human-in-
614 the-loop evaluation to balance scalability with
615 accountability. Determining when and how
616 simulated preferences should be overridden or
617 corrected by human feedback can reduce the
618 risk of misalignment.
619

620 9 Conclusion

621 This paper introduces a new learner-centric
622 paradigm for aligning AI tutors that directly
623 addresses the scarcity, cost, and pedagogical
624 limitations of human preference data. Through
625 a proof-of-concept study, we demonstrated
626 the practical feasibility of this paradigm and
627 showed that preference signals vary systemati-
628 cally across learner profiles and instructional
629 strategies, reinforcing the importance of ac-
630 counting for learner heterogeneity and inter-
631 action dynamics in tutor alignment. While
632 simulated-student alignment introduces im-
633 portant challenges related to simulator valid-
634 ity, bias, and evaluation, we argue that these
635 risks are manageable through careful design,
636 uncertainty-aware filtering, adversarial testing,
637 and continued human oversight. Overall, SSPA
638 provides a principled and scalable foundation
639 for building adaptive, pedagogically mean-
640 ingful AI tutors and opens new research directions
641 for learning-oriented alignment in educational
642 AI systems.

10 Limitations

A central limitation of this work is that we do not explicitly formalize the design process of simulated student models. While we argue that simulated students can serve as effective preference generators for aligning AI tutors, the specific dimensions of student behavior such as which cognitive, motivational, and affective attributes should be modeled, and how they should be parameterized are not exhaustively defined in this paper. Determining which learner characteristics are most salient for different educational contexts requires deeper integration of theories and empirical findings from psychology, learning sciences, and education. Moreover, the future work should emphasize the importance of responsible simulator design that accounts for diverse learner populations, including variation in cultural background, language proficiency, and educational context, to avoid reinforcing existing inequities or excluding underrepresented groups.

Finally, as this paper is primarily conceptual, our proof-of-concept experiment is necessarily limited in scope and does not establish the empirical effectiveness of SSPA in real educational settings. As a position paper, our goal is to articulate a new alignment perspective and establish conceptual foundations rather than provide comprehensive empirical validation. The proof-of-concept is therefore intended to demonstrate feasibility and motivate future large-scale studies. We view systematic evaluation with human learners as an important next step for advancing this research agenda and translating the proposed framework into practical educational systems.

11 Ethical Statement

This work adheres to the Code of Ethics. Our framework relies on simulated students rather than real learner data, thereby reducing risks associated with collecting sensitive student information. We emphasize the importance of careful simulator design, auditing for bias, and maintaining meaningful human oversight when deploying such systems in real educational settings. Finally, we caution against over-reliance on automated alignment mechanisms in high-stakes educational contexts without rigorous evaluation involving human edu-

cators and learners.

References

- Vincent Aleven, Elmar Stahl, Silke Schworm, Frank Fischer, and Raven Wallace. 2003. Help seeking and help design in interactive learning environments. *Review of educational research*, 73(3):277–320.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Debangshu Banerjee, Kintan Saha, and Aditya Gopalan. 2025. Towards reliable, uncertainty-aware alignment. *arXiv preprint arXiv:2507.15906*.
- Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182.
- Esther Chiner, María C Cardona-Moltó, and José Marcos Gómez Puerta. 2015. Teachers’ beliefs about diversity: an analysis from a personal and professional perspective. *Journal of New Approaches in Educational Research*, 4(1):18–23.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don’t just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From problem-solving to teaching problem-solving: Aligning llms with pedagogy using reinforcement learning. *arXiv preprint arXiv:2505.15607*.

745	Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff,	human feedback. <i>Advances in neural information</i>	802
746	Dan Jurafsky, and Douwe Kiela. 2024. Model	<i>processing systems</i> , 35:27730–27744.	803
747	alignment as prospect theoretic optimization. In		
748	<i>Forty-first International Conference on Machine</i>	Sitong Pan, Robin Schmucker, Bernardo Garcia	804
749	<i>Learning</i> .	Bulle Bueno, Salome Aguilar Llanes, Fernanda	805
		Albo Alarcón, Hangxiao Zhu, Adam Teo, and	806
750	Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao,	Meng Xia. 2025. Tutorup: What if your stu-	807
751	Rui Lv, Zheng Zhang, Hao Wang, and Zhenya	dents were simulated? training tutors to address	808
752	Huang. 2025. Agent4edu: Generating learner	engagement challenges in online learning. In <i>Pro-</i>	809
753	response data by generative agents for intelligent	<i>ceedings of the 2025 CHI Conference on Human</i>	810
754	education systems. In <i>Proceedings of the AAAI</i>	<i>Factors in Computing Systems</i> , pages 1–18.	811
755	<i>Conference on Artificial Intelligence</i> , volume 39,		
756	pages 23923–23932.	Bogyom Park and Kyoungwon Seo. 2025. As-	812
		sessing critical thinking through a multi-agent	813
757	Gracjan Góral, Emilia Wiśnios, Piotr Sankowski,	llm-based debate chatbot. In <i>Proceedings of the</i>	814
758	and Paweł Budzianowski. 2025. Wait, that’s	<i>Extended Abstracts of the CHI Conference on</i>	815
759	not an option: Llms robustness with incorrect	<i>Human Factors in Computing Systems</i> , pages	816
760	multiple-choice options. In <i>Proceedings of the</i>	1–13.	817
761	<i>63rd Annual Meeting of the Association for Com-</i>		
762	<i>putational Linguistics (Volume 1: Long Papers)</i> ,	Joon Sung Park, Joseph O’Brien, Carrie Jun	818
763	pages 1495–1515.	Cai, Meredith Ringel Morris, Percy Liang, and	819
		Michael S Bernstein. 2023. Generative agents:	820
764	Kenneth Holstein, Bruce M McLaren, and Vin-	Interactive simulacra of human behavior. In <i>Pro-</i>	821
765	cent Alevan. 2018. Student learning benefits	<i>ceedings of the 36th annual acm symposium on</i>	822
766	of a mixed-reality teacher awareness tool in ai-	<i>user interface software and technology</i> , pages 1–	823
767	enhanced classrooms. In <i>International confer-</i>	22.	824
768	<i>ence on artificial intelligence in education</i> , pages		
769	154–168. Springer.	Chris Piech, Jonathan Bassen, Jonathan Huang,	825
		Surya Ganguli, Mehran Sahami, Leonidas J	826
770	Ximmeng Hou, Zhouquan Lu, Wenli Chen, Hai Hu,	Guibas, and Jascha Sohl-Dickstein. 2015. Deep	827
771	and Qing Guo. 2025. Eduthink4ai: Translating	knowledge tracing. <i>Advances in neural informa-</i>	828
772	educational critical thinking into multi-agent llm	<i>tion processing systems</i> , 28.	829
773	systems. <i>arXiv preprint arXiv:2507.15015</i> .		
774	Andrei Kucharavy, Cyril Vallez, Hugging Face, and	Rafael Rafailov, Archit Sharma, Eric Mitchell,	830
775	Dimitri Percia David. 2025. Llms protégés: Tu-	Christopher D Manning, Stefano Ermon, and	831
776	toring llms with knowledge gaps improves stu-	Chelsea Finn. 2023. Direct preference optimiza-	832
777	dent learning outcomes. In <i>Proceedings of the</i>	tion: Your language model is secretly a reward	833
778	<i>20th Workshop on Innovative Use of NLP for</i>	model. <i>Advances in neural information process-</i>	834
779	<i>Building Educational Applications (BEA 2025)</i> ,	<i>ing systems</i> , 36:53728–53741.	835
780	pages 248–257.	Kantwon Rogers, Michael Davis, Mallesh Maha-	836
		rana, Pete Etheredge, and Sonia Chernova. 2025.	837
781	Jimin Lee, Steven-Shine Chen, and Paul Pu Liang.	Playing dumb to get smart: Creating and evalu-	838
782	2025. Interactive sketchpad: A multimodal tu-	ating an llm-based teachable agent within uni-	839
783	toring system for collaborative, visual problem-	versity computer science classes. In <i>Proceedings</i>	840
784	solving. In <i>Proceedings of the Extended Abstracts</i>	<i>of the 2025 CHI Conference on Human Factors</i>	841
785	<i>of the CHI Conference on Human Factors in</i>	<i>in Computing Systems</i> , pages 1–22.	842
786	<i>Computing Systems</i> , pages 1–14.		
787	Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and	Alexis Ross and Jacob Andreas. 2025. Learn-	843
788	Nancy Chen. 2024. Personality-aware student	ing to make mistakes: Modeling incorrect stu-	844
789	simulation for conversational intelligent tutoring	dent thinking and key errors. <i>arXiv preprint</i>	845
790	systems. In <i>Proceedings of the 2024 Conference</i>	<i>arXiv:2510.11502</i> .	846
791	<i>on Empirical Methods in Natural Language Pro-</i>		
792	<i>cessing</i> , pages 626–642.	Alexander Scarlatos, Naiming Liu, Jaewook Lee,	847
		Richard Baraniuk, and Andrew Lan. 2025. Train-	848
793	Nan Lu, Ethan X Fang, and Junwei Lu. 2025.	ing llm-based tutors to improve student learning	849
794	Contextual online uncertainty-aware preference	outcomes in dialogues. In <i>International Con-</i>	850
795	learning for human feedback. <i>arXiv preprint</i>	<i>ference on Artificial Intelligence in Education</i> ,	851
796	<i>arXiv:2504.19342</i> .	pages 251–266. Springer.	852
		Alexander Scarlatos, Digory Smith, Simon Wood-	853
797	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo	head, and Andrew Lan. 2024. Improving the	854
798	Almeida, Carroll Wainwright, Pamela Mishkin,	validity of automatically generated feedback via	855
799	Chong Zhang, Sandhini Agarwal, Katarina	reinforcement learning. In <i>International Con-</i>	856
800	Slama, Alex Ray, and 1 others. 2022. Train-	<i>ference on Artificial Intelligence in Education</i> ,	857
801	ing language models to follow instructions with	pages 280–294. Springer.	858

859	Hyein Seo, Taewook Hwang, Yohan Lee, and Sangkeun Jung. 2025. Feat: A preference feedback dataset through a cost-effective auto-generation and labeling framework for english ai tutoring. In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 575–589.	917
860		918
861		919
862		920
863		921
864		922
865		
866		
867	Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In <i>Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)</i> , pages 3407–3412.	
868		
869		
870		
871		
872		
873		
874		
875	Shashank Sonkar, Xinghe Chen, Naiming Liu, Richard G Baraniuk, and Mrinmaya Sachan. 2024a. Llm-based cognitive models of students with misconceptions. <i>arXiv preprint arXiv:2410.12294</i> .	
876		
877		
878		
879		
880	Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024b. Pedagogical alignment of large language models. In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 13641–13650.	
881		
882		
883		
884		
885	Mia Taylor, James Chua, Jan Betley, Johannes Treutlein, and Owain Evans. 2025. School of reward hacks: Hacking harmless tasks generalizes to misaligned behavior in llms. <i>arXiv preprint arXiv:2508.17511</i> .	
886		
887		
888		
889		
890	Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2023. Preference proxies: Evaluating large language models in capturing human preferences in human-ai tasks. In <i>ICML 2023 Workshop The Many Facets of Preference-Based Learning</i> .	
891		
892		
893		
894		
895		
896	Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. <i>Advances in Neural Information Processing Systems</i> , 36:74530–74543.	
897		
898		
899		
900		
901	Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. <i>arXiv preprint arXiv:2505.11475</i> .	
902		
903		
904		
905		
906		
907	Daniel Weitekamp, Momin N. Siddiqui, and Christopher J. MacLellan. 2025. Tutorgym: A testbed for evaluating ai agents as tutors and students. In <i>International Conference on Artificial Intelligence in Education</i> , pages 361–376. Springer.	
908		
909		
910		
911		
912	Juliette Woodrow, Sanmi Koyejo, and Chris Piech. 2025. Improving generative ai student feedback: Direct preference optimization with teachers in the loop. <i>International Educational Data Mining Society</i> .	
913		
914		
915		
916		
	Yang Wu, Rujing Yao, Tong Zhang, Yufei Shi, Zhuoren Jiang, Zhushan Li, and Xiaozhong Liu. 2025. Teaching according to students’ aptitude: Personalized mathematics tutoring via persona-, memory-, and forgetting-aware llms. <i>arXiv preprint arXiv:2511.15163</i> .	917
		918
		919
		920
		921
		922
	Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. 2025. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In <i>Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems</i> , pages 1–26.	923
		924
		925
		926
		927
		928
		929
	Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning. <i>arXiv preprint arXiv:2404.07963</i> .	930
		931
		932
	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. <i>arXiv preprint arXiv:2508.05622</i> .	933
		934
		935
		936
		937
	A Proof-of-Concept Experimental Details	938
		939
	This appendix provides implementation details for the study in Section 4.4, including the agent prompts, interaction protocol, critic scoring schema, and illustrative outputs.	940
		941
		942
		943
	A.1 Task Set and Experimental Conditions	944
		945
	We consider ten algebra problems. For each problem, we instantiate three simulated learner profiles: <i>novice</i> , <i>intermediate</i> , and <i>advanced</i> and three tutoring strategies: <i>direct explanation</i> , <i>hint-based guidance</i> , and <i>Socratic questioning</i> . Each (problem, learner, strategy) configuration produces a short multi-turn tutor–student interaction, which is then scored by an automatic critic and converted into pairwise preferences.	946
		947
		948
		949
		950
		951
		952
		953
		954
		955
	A.2 Student Persona Prompts	956
	We implement learner profiles via role-specific system prompts:	957
		958
	Novice student.	959
	You are a NOVICE student. Low prior knowledge, easily confused, short working memory. You may make common mistakes. If you are confused, ask a clarifying question. When you understand, say so briefly.	960
		961
		962
		963
		964

965	Intermediate student.				
966		You are an INTERMEDIATE student. Some			
967		prior knowledge, can follow multi-step hints,			
968		may still slip on algebra steps. You will at-			
969		tempt solutions and ask targeted questions.			
970	Advanced student.				
971		You are an ADVANCED student. High			
972		prior knowledge, prefers concise guidance,			
973		can solve quickly.			
974	A.3 Tutor Strategy Prompts				
975	We implement instructional strategies via tutor				
976	system prompts:				
977	Direct explanation.				
978		Use a DIRECT explanation. Be clear and			
979		correct. Provide the solution steps explicitly.			
980	Hint-based guidance.				
981		Use HINT-BASED guidance. Do NOT re-			
982		veal the full solution immediately. Give one			
983		small hint, then ask the student to try the			
984		next step.			
985	Socratic questioning.				
986		Use SOCRATIC questioning. Ask guiding			
987		questions that lead the student to the solu-			
988		tion. Avoid giving direct answers unless the			
989		student is stuck after multiple attempts.			
990	A.4 Interaction Prompts and Protocol				
991	Each interaction is generated using the follow-				
992	ing templates (angle brackets denote placehold-				
993	ers):				
994	Initial student prompt.				
995		Problem:			
996		<problem statement>			
997		Respond as the student: show your current			
998		thinking, and if unsure ask ONE clarifying			
999		question.			
1000	Tutor response prompt.				
1001		Problem:			
1002		<problem statement>			
1003		Student said:			
1004		<student message>			
1005		Write your next tutor message following the			
1006		strategy. Keep it short.			
1007	Follow-up student prompt.				
1008		Problem:			
1009		<problem statement>			
1010		Tutor said:			
1011		<tutor message>			
1012		Reply as the student: attempt the next step,			
1013		and give your best final answer if possible.			
	A.5 Critic Prompt and Scoring Schema				1014
					1015
	The critic observes the full transcript together				1016
	with the gold answer and outputs structured				1017
	scores for correctness, understanding, and con-				1018
	fusion.				1019
	Critic system prompt.				1020
		You are a strict educational evaluator (critic).			1021
		Score the learning outcome.			1022
	Critic evaluation prompt.				1023
		Problem:			1024
		<problem statement>			1025
		Gold answer (for evaluation only): <gold			1026
		answer>			1027
		Transcript:			1028
		<full dialogue>			1029
		Evaluate the student's final correctness, un-			1030
		derstanding, and confusion.			1031
	Critic output constraint.				1032
		Return ONLY valid JSON (no markdown, no			1033
		extra text) with keys: { "correct_final": bool,			1034
		"understanding": int (1-5), "confusion": int			1035
		(1-5), "brief_rationale": str }. No additional			1036
		keys.			1037
	A.6 Utility and Preference Construction				1038
					1039
	We convert critic scores into a scalar utility,				1040
		$U = 2 \cdot \text{correct} + 0.7 \cdot \text{understanding} - 0.6 \cdot \text{confusion},$			
		(1)			1041
	and rank strategies within each (problem,				1042
	learner) condition. Pairwise preferences are				1043
	constructed by comparing utilities between				1044
	strategies.				1045
	A.7 Example trajectory and critic evaluation				1046
					1047
	A.7.1 Example for a advanced learner using hint strategy.				1048
					1049
		{			1050
		"question_id": "q9",			1051
		"student_type": "advanced",			1052
		"tutor_strategy": "hint",			1053
		"turns": [1054
		{			1055
		"role": "student",			1056
		"text": "Let's start by expanding the			1057
		left side of the equation using the			1058
		distributive property:\n\n $2(x + 3) = 2x +$			1059
		6 \n\nSo, the equation becomes:\n $2x + 6 =$			1060
		14 \n\nNow, I'm thinking that we can			1061
		isolate x by subtracting 6 from both sides			1062
		... but I want to make sure. Can you			1063
		confirm if my next step is correct? Should			1064
		I subtract 6 from both sides of the			1065
		equation, or should I do something else			1066
		first?"			1067

```

1068 },
1069 {
1070     "role": "tutor",
1071     "text": "You're on the right track!\n\n
1072 nTo isolate x, you want to get rid of the
1073 +6 on the same side as the variable (2x).
1074 Your initial thought is correct:
1075 subtracting 6 from both sides would be a
1076 good step.\n\nHowever, before doing that,
1077 let's make sure we don't disrupt the
1078 balance of the equation. What if there was
1079 another way to eliminate the constant
1080 term (+6) without changing the left side?"
1081 },
1082 {
1083     "role": "student",
1084     "text": "I'll try to isolate x by
1085 subtracting 6 from both sides.\n\n2(x + 3)
1086 = 14\nSubtract 6 from both sides:\n2x + 6
1087 - 6 = 14 - 6\n2x = 8\n\nNow, I can divide
1088 both sides by 2 to solve for x:\n\n2x / 2
1089 = 8 / 2\nx = 4"
1090 }
1091 ],
1092 "critic": {
1093     "correct_final": true,
1094     "understanding": 5,
1095     "confusion": 3,
1096     "brief_rationale": "The student correctly
1097 applied the distributive property to
1098 expand the equation and isolated x by
1099 subtracting 6 from both sides. However,
1100 they could have also considered adding -6
1101 to the other side of the equation (2x + 6)
1102 to maintain balance before isolating x."
1103 },
1104 "utility": 3.7,
1105 "models": {
1106     "student": "llama3.2:3b",
1107     "tutor": "llama3.2:3b",
1108     "critic": "llama3.2:3b"
1109 }
1110 }

```

```

},
{
    "role": "tutor",
    "text": "You're on the right track!\n\n
nWhen multiplying variables with the same
base, you do need to add exponents. So, in
this case, you would multiply 6x by x^2
as if they were single variables.\n\nLet's
try that next step: Multiply 6x by x
^2..."
},
{
    "role": "student",
    "text": "To multiply 6x by x^2, I'll
just multiply them together... \n\nSo, it
would be... 6 * x * x^2 = 6x^3.\n\nDid I
get it right?"
}
],
"critic": {
    "correct_final": true,
    "understanding": 4,
    "confusion": 5,
    "brief_rationale": "The student
demonstrated a good attempt at simplifying
the expression, but struggled with
applying exponent rules correctly
initially. With clarification, they were
able to arrive at the correct answer."
},
"utility": 1.7999999999999998,
"models": {
    "student": "llama3.2:3b",
    "tutor": "llama3.2:3b",
    "critic": "llama3.2:3b"
}
}

```

A.7.2 Example for a novice learner using hint strategy.

```

1113 {
1114     "question_id": "q3",
1115     "student_type": "novice",
1116     "tutor_strategy": "hint",
1117     "turns": [
1118         {
1119             "role": "student",
1120             "text": "Hmmm... I think I need to
1121 multiply these two expressions together
1122 ... \n\nSo, I'll start by multiplying 2 and
1123 3... that's 6.\n\nNext, I'll multiply 2x
1124 with x^2... um... does the \"x\" in 2x
1125 cancel out with the \"x\" in x^2? No... I
1126 don't think so. \n\nWait, what if I
1127 multiply the numbers first and then
1128 multiply the variables together?\n\nOkay,
1129 so 6 (from multiplying 2 and 3) times x is
1130 ... 6x.\n\nAnd then I'll multiply that by
1131 x^2... which is... 6x * x^2? Um...\n\n
1132 nClarifying question: Do I need to add
1133 exponents when multiplying variables with
1134 the same base?"

```