

DYSAS: Dynamic State-Aware Student Simulation Beyond Static Personas

Partha Sarathi Purkayastha

Department of Computer Science
ETH Zürich
Zürich, Switzerland
ppurkayastha@ethz.ch

Heejin Do

Department of Computer Science
ETH Zürich
ETH AI Center
Zürich, Switzerland
heejin.do@ai.ethz.ch

Abstract

Simulated students play a central role in training and evaluating AI tutor agents, yet most approaches rely on static personas or unconstrained generation and implicitly assume stationary behavior. Real student responses also depend on transient learner states such as affect, attention, and memory that evolve during interaction. We introduce DYSAS, a dynamic state-aware student-simulation framework with three tutoring-facing modules, *Affect (A)*, *Attention (T)*, and *Memory (M)*. Their low, medium, and high levels form the discrete ATM27 state space. We develop this interface through Dynamic Learner States (DLS), a prefix-only representation inferred at each turn without future leakage and used to condition student generation. We evaluate static and dynamic variants on the Eedi benchmark across prompting, supervised fine-tuning, and preference optimization. No configuration wins every objective as each of correctness-state match, misconception behavior, social tone, lexical fidelity, and tutor-response compatibility favors different settings. Thus, we present DYSAS as an audit mechanism within a multi-objective simulator-selection framework, not as a universally superior add-on.

1 Introduction

AI tutoring systems increasingly rely on Large Language Models (LLMs) to provide scalable and interactive learning support (Shi et al., 2026; Scarlatos et al., 2025b; Dinucu-Jianu et al., 2025). A central challenge in developing and evaluating such systems is limited access to real students, which makes controlled experimentation expensive, slow, and difficult to reproduce (Scarlatos et al., 2026; Weitekamp et al., 2025; Martynova et al., 2025). Many approaches therefore use simulated students to train, stress-test, and benchmark tutoring systems (Scarlatos et al., 2026; Weitekamp et al., 2025; Martynova et al., 2025).

From an agent-evaluation perspective, the simulated student is the interactive environment against which a tutor agent is selected. Therefore, reliability requires the environment state at a turn t to depend only on information available by turn t . Conditioning on later student behavior or tutor scaffolds can otherwise make an offline tutor evaluation optimistically biased.

Recent work questions the validity of these simulators (Yuan et al., 2026; Scarlatos et al., 2026; Martynova et al., 2025). They can produce plausible surface behavior while missing the dynamics of human learning, a gap described as the *competence paradox* (Yuan et al., 2026). We argue that this limitation stems partly from a mismatch between the conditioning interface and the phenomenon being simulated. Existing approaches often rely on static personas or unconstrained generation (Scarlatos et al., 2026; McCrae and John, 1992; Goldberg, 1990). A common persona baseline is OCEAN, the Big Five model comprising Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Because OCEAN provides a compact and interpretable account of stable social and behavioral tendencies (Goldberg, 1990; McCrae and John, 1992; John and Srivastava, 1999; Scarlatos et al., 2026), it is a strong baseline for modeling social style. However, it was not designed to represent learners’ evolving cognitive and attentional states, e.g., whether they are currently confused, distracted, or unable to reuse a recent scaffold.

Learning is inherently dynamic (D’Mello and Graesser, 2012; Yuan et al., 2026). A learner’s response reflects both stable traits (McCrae and John, 1992; Goldberg, 1990) and transient states such as affect (Pekrun, 2006; D’Mello and Graesser, 2012), attention (Posner and Petersen, 1990), and working memory (Baddeley, 1992). These states evolve throughout interaction. Ignoring them can yield simulations that appear plausible yet fail to capture how students learn and change via instruction over

time (Yuan et al., 2026; Martynova et al., 2025). Student simulation is therefore partly a problem of modeling *state dynamics*, not merely surface realism (Yuan et al., 2026).

We propose **Dynamic State-Aware Student Simulation (DYSAS)** as a shift from static persona conditioning to explicit learner-state control. DYSAS uses *Affect, Attention, and Memory*, where each axis has low, medium, and high levels for varying learner-state configurations. We call this compact control space **ATM27**, where A, T, and M denote Affect, Attention, and Memory and 27 denotes their $3 \times 3 \times 3$ joint configurations. It is an operational interface for tutoring dialogue, not a psychometric diagnosis. The contrast with OCEAN is functional rather than ideological: OCEAN represents broad social disposition and interaction style, while ATM27 represents the short-horizon learning condition that can change the tutor’s next action.

We make the interface dynamic through **Dynamic Learner States (DLS)**. At each student turn, DLS infers a state from the available dialogue prefix and conditions generation on that state. The current state never observes future student behavior or future tutor scaffolds. This prefix-only rule removes avoidable leakage and turns a hidden conditioning assumption into an explicit trajectory that can be inspected turn by turn. Our evaluation predicts each held-out student turn from its authentic dialogue prefix, so it measures next-turn fidelity rather than the accumulation of errors in a free-running interaction. We evaluate DLS and retrospective full-dialogue static comparators for both OCEAN and ATM27 across prompting, supervised fine-tuning (SFT), and Direct Preference Optimization (DPO) on the Eedi student-simulation benchmark (Scarlatos et al., 2026).

Contributions. We make three contributions. First, we introduce ATM27, a tutoring-native learner-state interface organized as Affect, Attention, and Memory. Second, we instantiate dynamic conditioning through DLS, a prefix-only update mechanism that applies across prompting, SFT, and DPO and exposes auditable state trajectories. Third, our controlled grid shows that evaluation objectives select different simulator configurations; we turn these trade-offs into a goal-to-configuration map instead of collapsing them into one leaderboard score.

2 Related Work

2.1 Simulated students for AI tutoring

LLM-based simulated students support tutor evaluation and teacher-facing practice, but recent work has made their limitations unusually visible. Prior work simulates students at different skill levels (Benedetto et al., 2024), while Wu et al. (2025) target diverse cognitive levels and show that helpful-assistant priors can produce overly advanced answers that suppress realistic mistakes. CIMA (Stasaski et al., 2020), MathDial (Macina et al., 2023), PIIvot (Zent et al., 2025), TutorGym (Weitekamp et al., 2025), and the Eedi/QATD2K benchmark (Scarlatos et al., 2026) show that prompting underperforms post-training on several realism dimensions and that strong systems remain far from real student behavior. Teacher-facing studies also find students that are overly polished, overly attentive, emotionally flat, or logically inconsistent (Martynova et al., 2025; Cao et al., 2027). TutorGym and classroom-scale simulators demonstrate practical value while emphasizing that interaction trajectories matter more than isolated answers (Weitekamp et al., 2025; Xu et al., 2025). Complementary tutor-side work uses budgeted verification to improve next-action reliability (Purkayastha, 2027), whereas our focus is the fidelity of the simulated learner used for tutor evaluation. Recent validity work argues that student simulation should be judged by epistemic fidelity rather than surface realism alone (Yuan et al., 2026; Song et al., 2026; Do et al., 2026). We focus on a missing component, an interpretable turn-level learner-state interface with a prefix-only update rule.

2.2 Learner modeling beyond generic personas

In intelligent tutoring systems, the learner model enables adaptation rather than serving as decorative metadata. Recent history-conditioned student simulation summarizes longitudinal question-answering and dialogue records into learner profiles (Duan et al., 2026). Classical work models knowledge acquisition, forgetting, and uncertainty over time (Corbett and Anderson, 1995; Pavlik et al., 2009; Piech et al., 2015; Abdelrahman et al., 2023; Yang et al., 2026). Other work studies affect, self-regulation, and open learner models that expose learner state for pedagogical intervention (Pekrun, 2006; Winne and Hadwin, 1998; Zimmerman, 2002; Bull and Kay, 2007, 2010; Xhakaj and

Verbert, 2021; Robles Mucho et al., 2025). These traditions motivate local, updateable, and pedagogically actionable descriptors. OCEAN traits instead describe broad stable tendencies (Goldberg, 1990; McCrae and John, 1992; McCrae and Costa, 1997; John and Srivastava, 1999). They remain a useful baseline, but do not indicate whether a student has momentarily lost the thread of a scaffold. Thus, we compare against OCEAN while centering ATM27 on the turn-local state that conditions the next simulated response.

2.3 Conditioning and post-training for student simulation

Persona conditioning is common practice in dialogue generation, but its benefits are often modest and task-dependent. Early persona-dialogue models focused on style consistency (Li et al., 2016; Zhang et al., 2018). More recent studies find that persona variables explain only a limited fraction of human variation and degrade over long interactions (Hu and Collier, 2024; Luz de Araujo et al., 2026; Ni et al., 2026). Unlike history-conditioned profiles derived from longitudinal records (Duan et al., 2026), ATM27 is a discrete, prefix-only turn state updated from the current dialogue prefix. In education, prompting, SFT, and DPO have all been used to steer tutors or learners, but each changes *how* conditioning information is consumed. Prompting can surface high-level instructions, SFT can internalize recurring patterns, and preference optimization can further reshape behavior toward task-specific desiderata (Sonkar et al., 2024; Puech et al., 2025; Rafailov et al., 2023; Lee et al., 2026). Our question is therefore not merely whether learner-state conditioning helps, but when a given representation remains useful after post-training.

3 Method

3.1 Task and setting

We study turn-level student simulation in tutoring dialogues. Let h_t be the dialogue prefix ending at the tutor turn before the next student response, and let y_t be the simulated next student turn. A simulator produces

$$y_t \sim p_\theta(\cdot \mid h_t, c_t), \quad (1)$$

where c_t is the conditioning signal. It can be absent or can encode a static or dynamic OCEAN or ATM27 profile. Evaluation compares the generated turn with the held-out real turn using the

benchmark’s multi-metric suite. All conditions share the generator family, training workflow, evaluation rows, and judge protocol. The retrospective static baselines infer one descriptor from the full dialogue, whereas DLS infers a state from the available prefix at each turn. Because this contrast changes both update frequency and information access, it compares complete conditioning regimes rather than identifying the causal effect of updating alone. Figure 1 shows the complete pipeline.

3.2 ATM27: Tutoring-Actionable Learner State

ATM27 represents the student state along three distinct dimensions, **Affect (A)**, **Attention (T)**, and **Memory (M)**, each rated on low, medium, or high levels. Their Cartesian product defines 27 states. Rather than aiming to capture all learner characteristics, we focus on these dimensions because they satisfy three desirable properties for tutoring-oriented state representations: they are transient, pedagogically actionable, and observable from dialogue. They evolve at the timescale of tutoring interactions, have well-established effects on a student’s immediate learning behavior, and can be inferred without latent psychological diagnosis (Section C.3).

Specifically, affect shapes persistence and help-seeking (Pekrun, 2006; Baker et al., 2010; D’Mello and Graesser, 2012; Loderer et al., 2020; Pekrun, 2024; Ryan et al., 2001), attention governs instructional uptake (Kahneman, 1973; Posner and Petersen, 1990; Schultz and Arroyo, 2014; Wong et al., 2022; Szpunar et al., 2013), and working memory supports the retention and reuse of recent scaffolds (Atkinson and Shiffrin, 1968; Baddeley, 1992, 2003; Sweller, 1988; Mayer, 2009; Peng et al., 2016; Paas and van Merriënboer, 2020). We therefore present ATM27 as an operational interface for tutoring, rather than a comprehensive clinical model of cognition or a psychometric diagnosis. Under this formulation, *Affect* captures local emotional stance such as anxiety, confidence, or willingness to continue. *Attention* captures whether the student is tracking the current tutoring move. *Memory* captures whether recent scaffolds and intermediate results remain available at the next turn. The conservative annotation rule assigns low or high only when multiple cues support it and otherwise defaults to medium.

ATM27 and OCEAN answer different control questions. OCEAN asks what a student is gen-

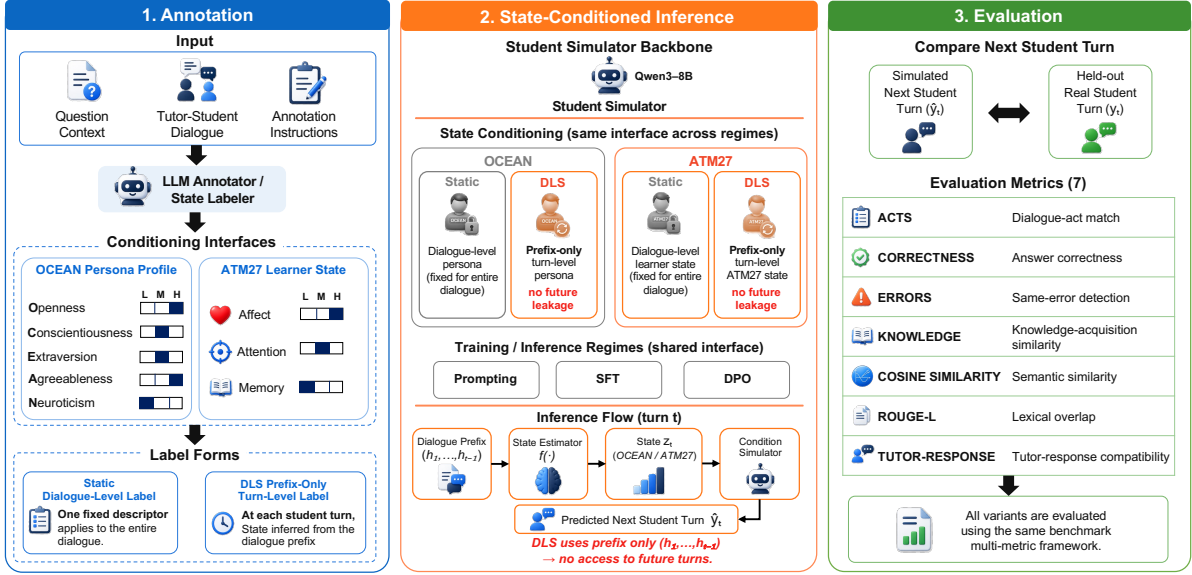


Figure 1: **Overview of DYSAS.** The framework separates two dimensions of student simulation. Student representation determines what is modeled, using either OCEAN traits or the tutoring-oriented ATM27 state, while student modeling determines how this condition evolves, using either one static label or prefix-only DLS updates at each turn. The workflow proceeds from annotation to state-conditioned next-turn generation and multi-metric evaluation across prompting, SFT, and DPO, with DLS preventing future leakage.

erally like. ATM27 asks what learning state the student appears to be in now. Low affect can yield hesitant or disengaged replies, low attention can lose the thread, and low memory can prevent reuse of a recent scaffold. Medium states represent partial resolution and high states represent confident, focused continuation. Neither interface is treated as ground truth or as a universal replacement for the other.

3.3 Dynamic learner states (DLS)

Static conditioning assigns one descriptor to the whole dialogue. DLS recomputes the state from the available prefix at every student turn. Let g be a state annotator and z_t the resulting state.

$$z_t = g(h_t), \quad y_t \sim p_\theta(\cdot | h_t, z_t). \quad (2)$$

The key restriction is causal. The annotator observes h_t and never future turns. Full-dialogue descriptors can otherwise leak future student behavior or tutor scaffolds into current-turn generation. The prefix-only path in Figure 1 makes this boundary explicit.

Operationally, a structured annotator maps the prefix to discrete labels. The labels are rendered as a short state block before the next student turn is generated. The same mechanism is used for ATM27 and OCEAN, so the comparison isolates

representation rather than mechanics. This design is consistent with turn-wise state tracing, knowledge tracing, verification, and open learner models (Scarlato et al., 2025a; Wang et al., 2025; Puech et al., 2025; Xhakaj and Verbert, 2021; Robles-Mucho et al., 2025). The label sequence proves that the explicit control state changes. It does not prove recovery of a student’s latent cognitive state. Because DLS adds neither a recurrent planner nor a separate simulator, it remains portable across prompting, SFT, and DPO. It also makes failures inspectable at three points, state inference, state rendering, and the generator’s use of the state.

3.4 Integration across Prompt, SFT, and DPO

Prompting inserts the state block directly into the model context. SFT keeps the benchmark workflow unchanged except for an optional static or dynamic state block in the input. DPO uses the same conditioning structure in the preference prompt. The shared interface lets us test whether a representation helps before post-training, whether it remains useful afterward, and whether turn-level updates become redundant or destabilizing once the generator has internalized related patterns. Templates, insertion rules, and training details are provided in Section B.

4 Experimental Setup

4.1 Dataset and Tasks

We evaluate on the Eedi/QATD2K tutoring-dialogue benchmark (Eedi, 2025; Zent et al., 2025; Scarlatos et al., 2026). QATD2K contains anonymized English math tutoring dialogues grounded in real tutoring interactions, making it a useful test bed for student simulation. Following the benchmark’s turn-level dialogue assembly, each example consists of a dialogue prefix ending at a tutor turn, and the simulator must generate the next student turn. We therefore study *turn-level* student simulation rather than full-dialogue generation.

The resulting train/validation/test split contains 1,147/382/382 dialogues and 13,079/4,439/4,225 student turns, respectively. All compared systems are evaluated on the same aligned 4,225-turn test set from 382 dialogues. Dynamic ATM27/OCEAN labels are constructed prefix-only. As explicitly noted by Scarlatos et al. (2026), their full-dialogue OCEAN baseline uses ground-truth information unavailable at the target turn. Following this setup, our reimplemented static OCEAN baseline—and the analogous static ATM27 variant—use descriptors inferred from the full dialogue. We therefore retain both only as retrospective offline comparators. The appendix reports the state distributions and annotation prompts.

Student generators use Qwen3-8B, and state annotations use Qwen3-4B-Instruct-2507 (Yang et al., 2025; Qwen Team, 2025b,a). We use the 8B generator model as a practical capacity compute trade-off for the full prompting/SFT/DPO grid, and holding the generator fixed also prevents backbone scale from confounding conditioning effects. We selected the 4B annotator model because pilot 8B outputs were more concentrated only in a few generic states across train, validation, and test (Section C.2). As diversity is not validity, we also conducted a coverage-aware, single-human audit of 100 static training dialogues. Human agreement is higher for Qwen3-4B-Instruct-2507 than Qwen3-8B on Affect/Attention/Memory ($\kappa = .710/.686/.737$ vs. $.300/.478/.322$) and on the full ATM27 state (.570 vs. .181). Raw agreement and sampling details are provided in Table 7. We keep the 4B annotator fixed across conditions, and the same structured state format is reused in prompting, SFT, and DPO on the 8B generator.

4.2 Metrics and Interpretation

We report the seven benchmark metrics of Scarlatos et al. (2026): ACTS, CORRECTNESS, ERRORS, KNOWLEDGE, COSINE SIMILARITY, ROUGE-L, and induced TUTOR-RESPONSE fit (Scarlatos et al., 2026; Lin, 2004; Reimers and Gurevych, 2019). The metric definitions are detailed in Table 4. All metrics are higher-is-better. ERRORS deserves special care as it measures whether the simulated student reproduces the reference student’s *error pattern* on turns where the real student is wrong. It is therefore a conditional realism metric, not an answer-quality penalty. The denominator also matters. On the aligned test set, 24 of 4,225 rows are skipped by the correctness and error judge path. Among the remaining 4,201 rows, 2,305 turns are usable for correctness-style evaluation, and only 908 incorrect-turn cases contribute to ERRORS. Because CORRECTNESS and ERRORS are only defined on subsets of the aligned test set, we interpret them jointly with the global similarity metrics rather than as standalone rankings.

4.3 Conditioning families and Training Protocol

We compare three training regimes. **Prompt** (ZERO-SHOT, ICL, OCEAN-STATIC, OCEAN-DLS, ATM27-STATIC, ATM27-DLS), **SFT** (no persona plus the four persona variants), and **DPO** with the same five post-trained variants. We use Qwen3-8B throughout and keep decoding, optimization logic, question context, target rows, and evaluation fixed. SFT uses LoRA-style adaptation (Hu et al., 2022; Dettmers et al., 2023), and DPO follows Rafailov et al. (2023) on the SFT checkpoints. Static variants reuse one rendered full-dialogue descriptor; dynamic variants append a turn-specific prefix-only state block. Thus, OCEAN versus ATM27 changes *what* is represented, while static versus DLS changes both update timing and the information boundary. Templates, insertion rules, and training details appear in the appendix.

On the evaluator side, the benchmark already distills GPT-4.1-derived labels into local Llama support models for dialogue acts and uses learned local models for knowledge and tutor-response scoring (Scarlatos et al., 2026). Its published headline CORRECTNESS/ERRORS path instead prompts GPT-5 mini. We extend the benchmark’s local-support-model recipe to these two metrics with fine-tuned

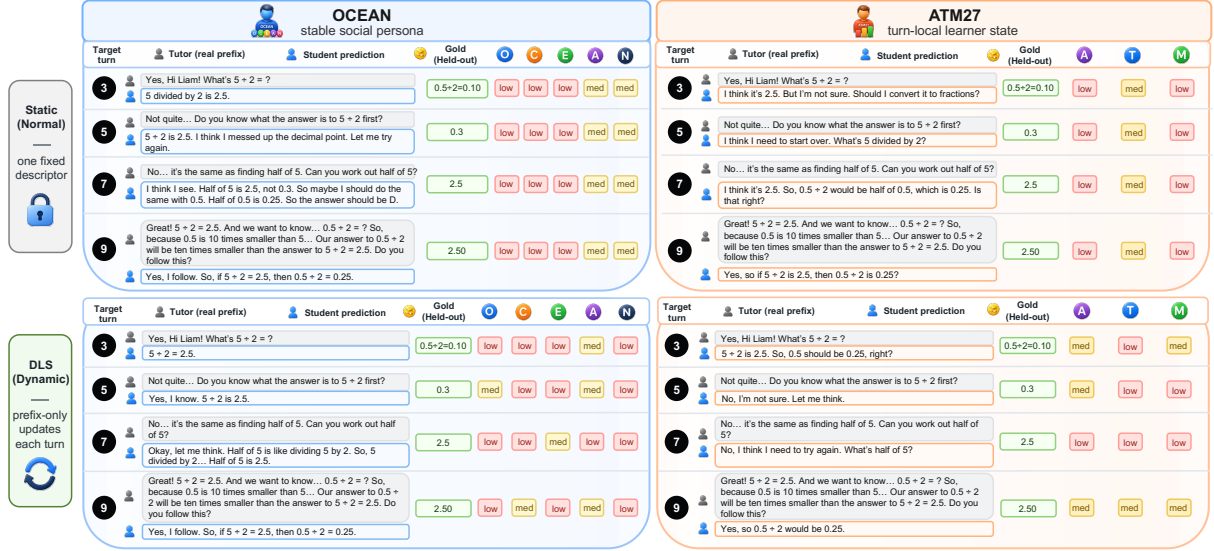


Figure 2: **Fixed-prefix next-turn comparison.** Every prediction uses the same held-out real prefix and the rows are not a free-running rollout. OCEAN versus ATM27 changes what is represented; Static versus DLS changes both update timing and whether the descriptor uses the full dialogue or only the available prefix.

meta-llama/Llama-3.1-8B-Instruct judges (Grattafiori et al., 2024), making every tested condition share the same evaluator checkpoints. The reported values therefore support controlled comparisons under one fixed judge protocol.

5 Results

Table 1 yields quantitative and qualitative findings on how training regime, representation, and update rule jointly shape simulated-student behavior. ATM27 and OCEAN emerge as complementary control interfaces rather than interchangeable persona labels.

5.1 Quantitative patterns

Case 1: Post-training raises the fidelity ceiling.

Across dialogue acts, knowledge, cosine similarity, ROUGE-L, and tutor-response compatibility, SFT and especially DPO outperform prompt-only generation. This agrees with prior student-simulation findings (Scarlatos et al., 2026; Martynova et al., 2025). Prompt models remain competitive on CORRECTNESS, but correctness alone can hide missing hesitation, misunderstanding, or help-seeking.

Case 2: Persona conditioning is not automatically better than no conditioning.

The post-trained NO-PERSONA baselines are strong. SFT NO-PERSONA is the best knowledge model and remains competitive on tutor compatibility. DPO NO-PERSONA stays near the top overall. Persona text helps only when it adds information the tuned gen-

erator has not internalized. Static ATM27 can add task-relevant local state, whereas broad OCEAN traits may become redundant or mismatched (Hu and Collier, 2024; Luz de Araujo et al., 2026; Yuan et al., 2026).

Case 3: ATM27 is the stronger static interface in this benchmark.

ATM27-STATIC beats OCEAN-STATIC on most prompt, SFT, and DPO metrics, most clearly after SFT. We interpret this as a representation effect. OCEAN supports broad social tone, while local confidence, focus, and recall are closer to the next tutoring response making even a coarse ATM27 descriptor more directly actionable (Desmarais and Baker, 2012; D’Mello and Graesser, 2012; Panadero, 2017; Yuan et al., 2026).

Case 4: DLS is regime-dependent, not a universal add-on.

In prompting, ATM27-DLS is the strongest controller because its state block directly guides whether the next reply is confident, distracted, stuck, or forgetful. The same update is less useful for OCEAN, whose traits are less tied to the immediate tutoring move. After post-training, the pattern partly reverses. OCEAN-DLS can rescue brittle OCEAN-STATIC behavior, whereas ATM27-DLS can over-constrain a generator that has internalized short-horizon ATM27-like cues (Li et al., 2016; Zhang et al., 2018; Rafailov et al., 2023; Yuan et al., 2026). We treat this as a mechanism hypothesis, rather than a causal conclusion, consistent with the transition evidence in

Table 1: Main Qwen3-8B student-simulation results using Llama-based evaluation judges. Best result within each stage is bolded.

Method	Acts	Corr.	Errors	Knowledge	Cos. Sim.	ROUGE-L	Tutor Resp.
<i>Prompting Methods</i>							
ZERO-SHOT	0.5308	0.8955	0.0562	0.7895	0.4728	0.1255	0.1731
ICL	0.4863	0.7223	0.0760	0.7739	0.4811	0.1051	0.1690
OCEAN-STATIC	0.5276	0.8898	0.0562	0.7905	0.4859	0.1301	0.1730
OCEAN-DLS	0.5164	0.8432	0.0507	0.7845	0.4560	0.1299	0.1749
ATM27-STATIC	0.5156	0.8411	0.1101	0.8046	0.5004	0.1383	0.1736
ATM27-DLS	0.5385	0.9012	0.0529	0.7991	0.5549	0.1856	0.1783
<i>Supervised Fine-Tuning (SFT)</i>							
NO-PERSONA	0.6126	0.7309	0.1553	0.8727	0.7123	0.2880	0.1886
OCEAN-STATIC	0.5437	0.5841	0.3348	0.8721	0.7097	0.2606	0.1876
OCEAN-DLS	0.6163	0.7738	0.1894	0.8684	0.7076	0.2872	0.1878
ATM27-STATIC	0.6185	0.7810	0.1806	0.8721	0.7138	0.2957	0.1879
ATM27-DLS	0.6061	0.7487	0.2048	0.8691	0.7053	0.2640	0.1888
<i>Direct Preference Optimization (DPO)</i>							
NO-PERSONA	0.6474	0.8289	0.0562	0.8760	0.7241	0.3056	0.1911
OCEAN-STATIC	0.6436	0.8676	0.0562	0.8710	0.7157	0.3039	0.1894
OCEAN-DLS	0.6479	0.8182	0.0507	0.8776	0.7245	0.3003	0.1915
ATM27-STATIC	0.6557	0.8654	0.0518	0.8773	0.7242	0.3175	0.1904
ATM27-DLS	0.6296	0.7810	0.0606	0.8770	0.7185	0.2855	0.1909

Table 2: **Use-case-specific simulator selection.** Each row reports the single configuration with the highest observed point estimate for that objective. These are objective-specific selections rather than a global ranking. Static configurations are retrospective full-dialogue comparators and should be used for offline analysis only.

Use case	Highest observed configuration	Evidence	Why use it
Help-seeking and uncertainty	Prompt ATM27-DLS	Uncertainty score: 2.149	Produces the strongest observed uncertainty and help-seeking behavior while using prefix-only dynamic states.
Correctness-state fidelity	Prompt ATM27-DLS	CORRECTNESS: .9012	Provides the highest match to the reference student’s answer-correctness state.
Misconception stress testing	SFT OCEAN-STATIC	ERRORS: .3348	Most strongly reproduces the reference student’s error family, although its lower CORRECTNESS (.5841) must be taken into consideration alongside this result.
Social-tone preservation	DPO ATM27-DLS	Social-tone score: 2.448	Achieves the highest observed social-tone composite on reference-social turns.
Dialogue-act matching	DPO ATM27-STATIC	ACTS: .6557	Best preserves the dialogue function of the reference student response.
Knowledge-acquisition fidelity	DPO OCEAN-DLS	KNOWLEDGE: .8776	Produces the highest match to the benchmark’s inferred student-knowledge state.
Semantic response similarity	DPO OCEAN-DLS	Cosine similarity: .7245	Provides the highest embedding-based semantic similarity to the reference student turn.
Lexical response fidelity	DPO ATM27-STATIC	ROUGE-L: .3175	Most closely preserves the reference response’s lexical content and sequence.
Tutor-response compatibility	DPO OCEAN-DLS	Tutor response: .1915	Most strongly induces a tutor response compatible with the response following the real student turn.

Section E.

5.2 Qualitative Patterns

Figure 2 shows how the four conditioning variants differ under a shared dialogue prefix. For the qualitative checks, we use transparent proxies for uncertainty, social tone, dynamic motion, and misconception-family errors. A separate human audit of the six proxy primitives on 100 turns found substantial agreement overall (macro/micro $\kappa = .610/.653$), although agreement was lower for the uncertainty (.551) and social (.520) cues that carry much of the qualitative interpretation. We therefore use these proxies as transparent surface diagnostics and anchor their aggregate patterns in

the examples provided in Section D, rather than treating lexical matches as semantic judgments.

Case 1: ATM27 captures local uncertainty better. On uncertain or help-seeking turns, ATM27 more often preserves compact, student-like uncertainty. It asks for help, reports confusion, or tentatively checks an answer, while OCEAN more often starts solving or expands into assistant-like explanation. A tutor should respond differently to “I do not know” than to a simulator that has performed part of the tutoring. Figure 4 and Table 13 show the clearest pattern in prompt and DPO settings. SFT compresses many replies into short templates and narrows the difference.

Case 2: OCEAN can preserve stable social tone better. Some student behavior reflects stable conversational style rather than local cognitive state. For greetings, thanks, polite closure, and cooperative acknowledgements, OCEAN often preserves the social move while ATM27 may continue calculation or become over-local. The interfaces are therefore complementary. ATM27 provides a turn-local representation of inferred learner state, and OCEAN preserves some stable social cues in our qualitative analysis. Figure 5 and Table 15 show the strongest OCEAN advantage in prompt and SFT settings. DPO with DLS is mixed because preference optimization shortens both families into compact acknowledgements.

6 Discussion

Objectives select the simulator. Table 2 is the practical contribution of the evaluation grid. Prompt ATM27-DLS is strongest for help-seeking and answer-correctness fidelity, SFT OCEAN-STATIC is most useful for misconception stress testing, and the DPO variants dominate several dialogue, knowledge, semantic, lexical, and tutor-response objectives. A collapsed average would hide these deployment-relevant choices, so simulator selection should follow the behavior being evaluated rather than an overall score.

DLS makes state motion auditable. In the aligned test split, ATM27-DLS visits all 27 ATM27 states and 62.4% of adjacent labels change, whereas static ATM27 covers 18 states. Table 19 links state changes to help-seeking, answer selection, apology, and closure. Even when ATM27-DLS drops after post-training, its trace localizes failure to state inference, rendering, or generation rather than hiding it inside one descriptor.

ERRORS must be read with fidelity. High ERRORS can identify misconception families for tutor tests, but also reward a simulator that is simply often wrong. So, we pair it with correctness and the other global similarity metrics for proper interpretation. Figure 6 and Table 21 show this contrast: SFT OCEAN-STATIC compromises correctness for same-error similarity, while prompt ATM27-DLS compromises error matching instead.

Training changes how side information is used. Prompted ATM27-DLS is directly action-guiding, while SFT can internalize short student patterns and make dynamic state text redundant or noisy. DPO

separates the objectives again. Side information can therefore be a scaffold in one regime and a distraction in another (Rafailov et al., 2023; Lee et al., 2026; Puech et al., 2025; Purkayastha, 2026). Determining whether this reflects internalization, redundancy, or noise requires targeted ablations.

Agent evaluation requires the right information boundary. The simulated student is part of the tutor agent’s environment. Retrospective static descriptors remain useful for offline diagnosis, but they are not causally available online - only DLS and no-persona rows (except ICL) satisfy that boundary here. A future hybrid could combine a slow profile with fast ATM27 states, provided both are inferred from past interaction only.

7 Future Work

Future work should extend ATM27 and proxy judgments to multiple educators and prevalence-sampled turns, incorporate multimodal evidence, and combine fast-changing learner states with slower interpersonal style. It should test transfer across subjects, benchmarks, simulator families, and simpler controls such as engagement or knowledge-state flags. Prefix-only traces could support slice audits and open learner models; the key test is whether they change tutor selection or policy outcomes under expert, classroom, and real-learner evaluation (Yuan et al., 2026; Scarlatos et al., 2026; Bull and Kay, 2010; Xhakaj and Verbert, 2021; Robles Mucho et al., 2025; Böck et al., 2025; Weitekamp et al., 2025).

8 Conclusion

DYSAS reframes simulated students as auditable components of tutor evaluation rather than static persona generators. On this benchmark, ATM27 is generally more task-aligned than OCEAN, while OCEAN better preserves some stable social behaviors. DLS adds prefix-only, turn-level state updates that make simulator behavior inspectable and avoid future leakage, but it does not produce uniform gains across prompting, SFT, DPO, or all evaluation metrics. The main result is therefore not one universally best simulator. It is a controlled framework for selecting and auditing simulators according to the educational behavior being tested. This supports clearer diagnosis, structured stress-testing, and safer offline iteration before evaluation with real learners (Scarlatos et al., 2026; Yuan et al., 2026; Weitekamp et al., 2025).

Limitations

We evaluate one tutoring benchmark, one dialogue domain, and one generator family, so the findings establish useful control abstractions rather than a cognitive theory. ATM27 compresses strategy, engagement, knowledge uncertainty, and other causes of similar surface behavior into discrete states and does not capture multimodal learner signals, which limits state identifiability. Different annotators also recover different normalized state distributions, so DYSAS treats inferred states as visible control assumptions rather than learner ground truth. Moreover, the human audit covers static full-dialogue ATM27 annotations rather than prefix-only DLS labels. Single-evaluator audits and multi-metric analysis support the reported within-benchmark comparisons, but automatic metrics and qualitative proxies cannot fully assess uncertainty, social appropriateness, or misconception fidelity. Thus, simulated students do not replace real learners, teacher judgment, or classroom evaluation (Scarlatos et al., 2026; Martynova et al., 2025; Yuan et al., 2026; Weitekamp et al., 2025), and broader generalization requires replication across domains, generator families, and real learning settings. The study therefore establishes a reproducible framework for selecting and diagnosing student simulators, with external validation as the next step.

Ethical Considerations

Leakage-free offline evaluation is a trustworthiness requirement. Simulated students do more than reduce experimental cost. They shape which tutors appear safe enough to advance to human studies. A simulator should therefore not use future turns to decide current learner behavior. Prefix-only DLS does not guarantee realism, but it removes an avoidable source of optimism and makes the conditioning signal causal with respect to the dialogue prefix; full-dialogue static descriptors do not satisfy this online boundary (Yuan et al., 2026; Scarlatos et al., 2026).

Student simulation should remain multi-objective. Prompt models can look strong on Correctness while remaining weak on Acts, language matching, and tutor-response compatibility. Post-trained systems can improve dialogue fidelity while losing some correctness. Safer offline tutor development should expose these trade-offs rather than hide them in one score. This is especially

important for ERRORS because it uses a smaller subset and can reward locally plausible but globally unresolved behavior.

Acknowledgments

This research was primarily supported by the ETH AI Center through an ETH AI Center postdoctoral fellowship to Heejin Do.

References

- Ghada Abdelrahman, Qing Wang, and Bernardete Ribeiro Nunes. 2023. [Knowledge tracing: A survey](#). *ACM Computing Surveys*, 55(11):1–37.
- Ron Artstein and Massimo Poesio. 2008. [Survey article: Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Richard C. Atkinson and Richard M. Shiffrin. 1968. [Human memory: A proposed system and its control processes](#). In *Psychology of Learning and Motivation*, volume 2, pages 89–195. Academic Press.
- Alan Baddeley. 1992. [Working memory](#). *Science*, 255(5044):556–559.
- Alan Baddeley. 2003. [Working memory: Looking back and looking forward](#). *Nature Reviews Neuroscience*, 4(10):829–839.
- Ryan S. J. d. Baker, Sidney K. D’Mello, Ma. Mercedes T. Rodrigo, and Arthur C. Graesser. 2010. [Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments](#). *International Journal of Human-Computer Studies*, 68(4):223–241.
- Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. [Using LLMs to simulate students’ responses to exam questions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, Miami, Florida, USA. Association for Computational Linguistics.
- Felix Böck, Michaela Ochs, Andreas Henrich, Dieter Landes, Jochen L. Leidner, and Yvonne Sedelmaier. 2025. [Learner models: Design, components, structure, and modelling](#). *User Modeling and User-Adapted Interaction*, 35(4):15.
- Susan Bull and Judy Kay. 2007. Student models that invite the learner in: The SMILI open learner modelling framework. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, pages 301–308.
- Susan Bull and Judy Kay. 2010. Open learner models. *International Journal of Artificial Intelligence in Education*, 20(1):1–39.

- Jie Cao, Ha Nguyen, Selim Yavuz, Boran Yu, Shuguang Wang, Pavneet Kaur Bharaj, and Dionne Cross Francis. 2027. [Developing authentic simulated learners for mathematics teacher learning: Insights from three approaches with large language models](#). In *Artificial Intelligence in Education*, pages 205–214, Cham. Springer Nature Switzerland.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Jacob Cohen. 1968. [Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit](#). *Psychological Bulletin*, 70(4):213–220.
- Albert T. Corbett and John R. Anderson. 1995. [Knowledge tracing: Modeling the acquisition of procedural knowledge](#). *User Modelling and User-Adapted Interaction*, 4(4):253–278.
- Michel C. Desmarais and Ryan S. Baker. 2012. [A review of recent advances in learner and skill modeling in intelligent learning environments](#). *User Modeling and User-Adapted Interaction*, 22(1–2):9–38.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*.
- 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](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 272–292, Suzhou, China. Association for Computational Linguistics.
- Sidney K. D’Mello and Arthur C. Graesser. 2012. [Dynamics of affective states during complex learning](#). *Learning and Instruction*, 22(2):145–157.
- Heejin Do, Shashank Sonkar, and Mrinmaya Sachan. 2026. Simulating students or sycophantic problem solving? on misconception faithfulness of llm simulators. *arXiv preprint arXiv:2605.12748*.
- Zhangqi Duan, Shuyan Huang, Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. 2026. [Who am i? history-aware profiles for student simulation in tutoring dialogues](#). *Preprint*, arXiv:2605.30051.
- Eedi. 2025. [Eedi: Question-anchored tutoring dialogues 2k \(QATD2K\)](#). Hugging Face Datasets. Accessed: 2026-05-08.
- Lewis R. Goldberg. 1990. [An alternative “description of personality”: The big-five factor structure](#). *Journal of Personality and Social Psychology*, 59(6):1216–1229.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10289–10307.
- Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, 2nd edition, pages 102–138. Guilford Press.
- Daniel Kahneman. 1973. *Attention and Effort*. Prentice-Hall.
- Jaewook Lee, Alexander Scarlatos, Simon Woodhead, and Andrew Lan. 2026. [Letting tutor personas speak up for LLMs: Learning steering vectors from dialogue via preference optimization](#). In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 78–92, San Diego, California, USA. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. [A persona-based neural conversation model](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.
- Kristina Loderer, Reinhard Pekrun, and James C. Lester. 2020. [Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments](#). *Learning and Instruction*, 70:101162.
- Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin

- Roth. 2026. [Persistent personas? role-playing, instruction following, and safety in extended interactions](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5329–5359, Rabat, Morocco. Association for Computational Linguistics.
- Jakub Macina, Nico Daheim, Sankalan Pal 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*.
- Daria Martynova, Jakub Macina, Nico Daheim, Özge Nilay Yalçın, Xiaoyu Zhang, and Mrinmaya Sachan. 2025. [Can LLMs effectively simulate human learners? teachers’ insights from tutoring LLM students](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 100–117.
- Richard E. Mayer. 2009. *Multimedia Learning*, 2nd edition. Cambridge University Press.
- Robert R. McCrae and Paul T. Costa. 1997. [Personality trait structure as a human universal](#). *American Psychologist*, 52(5):509–516.
- Robert R. McCrae and Oliver P. John. 1992. [An introduction to the five-factor model and its applications](#). *Journal of Personality*, 60(2):175–215.
- Bo Ni, Yu Wang, Leyao Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Luera, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen K. Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, and 11 others. 2026. [A survey on LLM-based conversational user simulation](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4266–4301, Rabat, Morocco. Association for Computational Linguistics.
- Fred Paas and Jeroen J. G. van Merriënboer. 2020. [Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks](#). *Current Directions in Psychological Science*, 29(4):394–398.
- Ernesto Panadero. 2017. [A review of self-regulated learning: Six models and four directions for research](#). *Frontiers in Psychology*, 8:422.
- Philip I. Pavlik, Jr., Hao Cen, and Kenneth R. Koedinger. 2009. [Performance factors analysis: A new alternative to knowledge tracing](#). In *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press.
- Reinhard Pekrun. 2006. [The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice](#). *Educational Psychology Review*, 18(4):315–341.
- Reinhard Pekrun. 2024. [Control-value theory: From achievement emotion to a general theory of human emotions](#). *Educational Psychology Review*, 36(3):83.
- Peng Peng, Jessica Namkung, Marcia Barnes, and Congying Sun. 2016. [A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics](#). *Journal of Educational Psychology*, 108(4):455–473.
- Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. [Deep knowledge tracing](#). In *Advances in Neural Information Processing Systems*.
- Michael I. Posner and Steven E. Petersen. 1990. [The attention system of the human brain](#). *Annual Review of Neuroscience*, 13:25–42.
- Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2025. [Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26291–26311.
- Partha Sarathi Purkayastha. 2026. [When thinking before retrieval hurts: TraceBound diagnostics for adaptive knowledge-graph retrieval](#). *Preprint*, arXiv:2607.24800. Accepted at the Failure Modes in Agentic AI (FAGEN) Workshop at ICML 2026.
- Partha Sarathi Purkayastha. 2027. [Reliability as a teammate: Budgeted verifier-in-the-loop \(BVIL\) policies for reliable LLM tutoring actions](#). In *Artificial Intelligence in Education*, pages 407–416, Cham. Springer Nature Switzerland.
- Qwen Team. 2025a. [Qwen/Qwen3-4B-Instruct-2507 model card](#). Hugging Face Model Hub. Accessed: 2026-05-08.
- Qwen Team. 2025b. [Qwen/Qwen3-8B model card](#). Hugging Face Model Hub. Accessed: 2026-05-08.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

- Jonathan L. Robles Mucho, Daniel C. Andrade-Girón, Violeta R. Benites-Tirado, Nilda B. Quispe-Maquera, Carmen Ayala-Jara, Hugo E. Tito Chura, Fernando M. Luna-Victoria, Karina M. Laura-De La Cruz, Orlando Rivera-Lozada, Ana A. Cerna Salcedo, Patricia S. Jaramillo Arica, and Juan J. Barboza. 2025. [Open learner models and pedagogical strategies in higher education: A meta-synthesis of approaches to self-regulated learning](#). *Frontiers in Education*, 10:1760183.
- Allison M. Ryan, Paul R. Pintrich, and Carol Midgley. 2001. [Avoiding seeking help in the classroom: Who and why?](#) *Educational Psychology Review*, 13(2):93–114.
- Alexander Scarlatos, Ryan S. Baker, and Andrew Lan. 2025a. [Exploring knowledge tracing in tutor-student dialogues using LLMs](#). In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 249–259.
- Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. 2026. [Simulated students in tutoring dialogues: Substance or illusion?](#) In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 42349–42385, San Diego, California, United States. Association for Computational Linguistics.
- Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025b. [Training LLM-based tutors to improve student learning outcomes in dialogues](#). In *Artificial Intelligence in Education*, pages 251–266, Cham. Springer Nature Switzerland.
- Sarah E. Schultz and Ivon Arroyo. 2014. [Tracing knowledge and engagement in parallel in an intelligent tutoring system](#). In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 312–315.
- Yuhong Shi, Kun Yu, Yifei Dong, and Fang Chen. 2026. [Large language models in education: A systematic review of empirical applications, benefits, and challenges](#). *Computers and Education: Artificial Intelligence*, 10:100529.
- Marina Sokolova and Guy Lapalme. 2009. [A systematic analysis of performance measures for classification tasks](#). *Information Processing & Management*, 45(4):427–437.
- Jiajia Song, Zhihan Guo, and Jionghao Lin. 2026. [Simulating novice students using machine unlearning and relearning in large language models](#). *Preprint*, arXiv:2603.26142.
- Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G. Baraniuk. 2024. [Pedagogical alignment of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
- Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. [CIMA: A large open access dialogue dataset for tutoring](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64.
- John Sweller. 1988. [Cognitive load during problem solving: Effects on learning](#). *Cognitive Science*, 12(2):257–285.
- Karl K. Szpunar, Novall Y. Khan, and Daniel L. Schacter. 2013. [Interpolated memory tests reduce mind wandering and improve learning of online lectures](#). *Proceedings of the National Academy of Sciences*, 110(16):6313–6317.
- Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. 2025. [Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of LLMs as your coding tutors](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12416–12436.
- Matthijs J. Warrens. 2012. [Cohen’s linearly weighted kappa is a weighted average](#). *Advances in Data Analysis and Classification*, 6(1):67–79.
- Daniel Weitekamp, Momin N. Siddiqui, and Christopher J. MacLellan. 2025. [TutorGym: A testbed for evaluating AI agents as tutors and students](#). In *Artificial Intelligence in Education*, pages 361–376, Cham. Springer Nature Switzerland.
- Philip H. Winne and Allyson F. Hadwin. 1998. [Studying as self-regulated learning](#). In Douglas J. Hacker, John Dunlosky, and Arthur C. Graesser, editors, *Metacognition in Educational Theory and Practice*, pages 277–304. Lawrence Erlbaum Associates.
- Aaron Y. Wong, Shelby L. Smith, Catherine A. McGrath, Lauren E. Flynn, and Caitlin Mills. 2022. [Task-unrelated thought during educational activities: A meta-analysis of its occurrence and relationship with learning](#). *Contemporary Educational Psychology*, 71:102098.
- Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. 2025. [Embracing imperfection: Simulating students with diverse cognitive levels using LLM-based agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9887–9908, Vienna, Austria. Association for Computational Linguistics.
- Fitim Xhakaj and Katrien Verbert. 2021. [Interactive learning dashboards and open learner models: A systematic review](#). *IEEE Transactions on Learning Technologies*, 14(6):643–657.
- 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 *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–26. ACM.

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.

Huali Yang, Junjie Hu, Shengze Hu, Zhuoran Xu, Xinjia Ou, Jing Geng, Linxia Tang, and Tao Huang. 2026. [A survey of deep learning based knowledge tracing from cognitive processing perspective](#). *Neurocomputing*, 660:131879.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. [FLASK: Fine-grained language model evaluation based on alignment skill sets](#). In *International Conference on Learning Representations*.

Zhihao Yuan, Yunze Xiao, Ming Li, Weihao Xuan, Richard Tong, Mona Diab, and Tom Mitchell. 2026. [Towards valid student simulation with large language models](#). *Preprint*, arXiv:2601.05473.

Matthew Zent, Digory Smith, and Simon Woodhead. 2025. [PIIvot: A lightweight NLP anonymization framework for question-anchored tutoring dialogues](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27479–27488, Suzhou, China. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*.

Barry J. Zimmerman. 2002. [Becoming a self-regulated learner: An overview](#). *Theory Into Practice*, 41(2):64–70.

A Appendix Overview

This appendix has five parts. Section B documents the conditioning families, data splits, prefix-only algorithm, and metric meanings. Section C.2 gives the complete shared 4B/8B annotation protocol, distribution study, production-model rationale, and single-human state-label audit. Section D expands the qualitative audit with proxy definitions, a separate single-human proxy validation, selected charts,

and compact mapped examples. Section E gives the state-motion analysis, including transition rates and a turn-level trace. Section F separates correctness from same-error realism and gives the corrected trade-off figure plus a table of exact readings.

B Experimental Protocol and Conditioning Families

B.1 Prefix-only dynamic conditioning

The central leakage-control rule for DLS is that the simulator may condition on the dialogue prefix only. For a target student turn, the state estimator receives the question context and real dialogue turns before that student turn, but it never sees the target response or later tutor scaffolding. The generated response is evaluated against the held-out real turn, and the next target is built independently from its real prefix rather than from a free-running rollout. The reimplemented static baselines are different: they use one full-dialogue descriptor and are retained only as retrospective offline comparators.

Line 4 is the only state-inference step. Because h_t stops before the target student turn, it enforces the information boundary that a deployed tutor would face. Each prediction uses its held-out real prefix; generated turns are not appended to later evaluation contexts. The algorithm is agnostic to whether the state vector is ATM27 or OCEAN.

B.2 Dataset splits and metric applicability

“Dialogues” counts unique tutoring conversations. “Student turns” counts the number of target student responses that can be generated from prefixes. “Evaluated turns” is the row count used for validation or test scoring. Correctness-style metrics are not defined for every row: in the aligned test set, 2,305 rows have a judgeable correctness target and 908 rows are incorrect-turn cases where ERRORS is defined. This is why the main paper treats ERRORS as conditional evidence rather than as a global quality score.

B.3 Metrics and what each score signifies

Table 4 explains why the main results are multi-objective. A simulator can score well on surface similarity while smoothing away uncertainty, or it can score well on ERRORS because it stays wrong too often. The qualitative audit exists to connect these aggregate scores back to observable student behaviors.

Algorithm 1 Prefix-only Dynamic Learner State (DLS) student simulation

Require: Dialogue dataset $\mathcal{D}$ with turn segmentation; state estimator g ; student model π .

```
1: for dialogue  $d \in \mathcal{D}$  do
2:   for student turn index  $t = 1, 2, \dots$  do
3:     Build prefix prompt  $h_t$  from the question context and all dialogue turns before student turn  $t$ .
4:     Infer state  $\hat{s}_t \leftarrow g(h_t)$ . {infer ATM27 or OCEAN from prefix only}
5:     Cache  $(d, t, \hat{s}_t)$  for the corresponding turn.
6:     Build generation prompt  $p_t \leftarrow \text{Format}(h_t, \hat{s}_t)$ .
7:     Generate simulated student turn  $\hat{u}_t \sim \pi(\cdot \mid p_t)$ .
8:     Score  $\hat{u}_t$  against held-out real turn  $u_t$ .
9:   end for
10: end for
```

Table 3: Aligned data split and metric applicability. Turn counts are student-turn examples used by the simulator.

Split	Dialogues	Student turns	Evaluated turns	Avg. student turns/dialogue
Train	1,147	13,079	–	11.40
Validation	382	4,439	4,439	11.62
Test	382	4,225	4,225	11.06

B.4 Training and decoding details

SFT and DPO follow the benchmark’s overall workflow. SFT is trained for three epochs with learning rate 5×10^{-5} , LoRA rank 32, and maximum input length 6000. DPO starts from the corresponding SFT checkpoint, uses $\beta = 0.1$, and optimizes preference triples built from the same state-conditioned prompt scaffold. Comparisons keep the generator family, question context, evaluation prefixes, target rows, and evaluator fixed. The side information and, for static versus DLS, its information boundary differ as documented above.

C Annotation Methodology

C.1 ATM27 Operational Semantics

The ATM27 rubric is grounded in achievement-emotion and affective-tutoring research (Pekrun, 2006; Baker et al., 2010; Pekrun, 2024), attention and engagement research (Posner and Petersen, 1990; Schultz and Arroyo, 2014), and working-memory and cognitive-load accounts (Baddeley, 1992, 2003; Sweller, 1988). These sources motivate the operational axes but do not establish the labels as latent psychological ground truth.

C.2 Shared 4B/8B Annotation Protocol and Selection

The 4B and 8B static files use the same ATM27 axis schema, have identical ordered dialogue keys, and preserve identical row-specific question/dialogue inputs for all 1,182/394/395

train/validation/test dialogues. Because raw serialization compliance could differ, we compare recovered axis values after the same normalization and ignore model-emitted state IDs. For ATM27, a normalized state is the ordered triple (Affect, Attention, Memory) with each value in {low, med, high}. The main pipeline uses the 4B annotator because its recovered states are broader and less concentrated than the 8B pilot outputs, and the human audit below provides a second comparison on matched dialogues.

In Figure 3, the left panel reports coverage: how many of the 27 possible ATM27 triples appear after normalization. The right panel reports concentration: how much mass is assigned to the single most frequent state. Lower coverage and higher top-state share mean the state signal is less diverse. Across train/validation/test, 4B recovers 22/20/19 states with entropy 2.61/2.65/2.68 and top-state shares 43.2%/40.9%/42.7%; 8B recovers 17/16/14 states with entropy 2.14/2.08/2.15 and top-state shares 62.2%/64.7%/59.5%. Thus, 8B is consistently more concentrated under the same protocol.

Table 6 gives the full numerical comparison. “Rows” is the number of raw dialogue-level annotations; “Norm.” is the subset with a recovered ordered triple. Entropy summarizes the full state distribution. The raw annotation audit precedes benchmark alignment, so its 395 test rows and 19 recovered 4B states differ from the aligned 382-dialogue evaluation, where static ATM27 covers

Table 4: Metric definitions used to interpret the main table. All metrics are higher-is-better, but they do not measure the same kind of realism.

Metric	What is compared	Interpretation and caveat
ACTS	Dialogue-act category of generated turn versus reference turn	Measures whether the simulator performs the same conversational function, such as answer, help request, acknowledgement, or off-topic move. High acts fidelity can still miss the exact content.
CORRECTNESS	Correctness state of generated answer versus reference correctness state	Captures whether the simulator is right or wrong in the same way at the answer-status level. It does not by itself tell whether a wrong answer comes from a realistic misconception.
ERRORS	Error pattern on rows where the reference student is wrong	Conditional same-error realism. Useful for misconception stress tests, but sparse and easy to misread if a model is broadly too wrong.
KNOWLEDGE	Knowledge-component / skill alignment of the generated response	Measures whether the response stays in the same math-skill neighborhood. This supports tutor-response realism but is not a substitute for correctness.
COSINE SIMILARITY	Sentence-embedding similarity between generated and reference turns	Captures semantic closeness. It can reward paraphrases but may overlook short social cues or small numerical errors.
ROUGE-L	Longest-subsequence lexical overlap	Captures surface wording similarity. It is helpful for compact student turns but should not dominate educational interpretation.
TUTOR RESPONSE	Compatibility of the tutor response induced by the generated student turn	Measures Inverse Perplexity score to determine whether the generated turn would lead to a tutor response similar to the reference continuation. This is the most downstream-facing metric in the table.

Table 5: Unified operational semantics and prefix-only inference rubric for ATM27. Each axis is defined through dialogue-facing control states rather than psychometric measurement. The cues are grounded in achievement emotions, attention, and working-memory/load theory (Pekrun, 2006; Posner and Petersen, 1990; Baddeley, 1992; Sweller, 1988).

Axis	Meaning	Low	Medium	High
Affect (A)	emotional confidence / engagement	anxious, low confidence (“I don’t know”, gives up, reassurance-seeking)	neutral, mildly uncertain (continues with occasional checking)	confident, persistent (decisive answers, self-corrects)
Attention (T)	task focus / tracking	distracted, off-task (answers wrong question, skips steps)	mostly on-task (minor drift, partial alignment)	fully focused (follows instruction precisely)
Memory (M)	recall of prior context	forgetful, inconsistent (contradictions, repeats questions)	partial recall (some reuse, occasional slips)	strong recall (consistent reuse, no contradictions)

18 states (Table 16). Broader coverage and lower concentration motivated the 4B production choice; these are control-signal properties rather than standalone evidence of label validity.

C.3 Single-human audit of static ATM27 labels

One human evaluator applied the same full-dialogue rubric to a stratified subset of 100 static training dialogues from the 866-dialogue population with valid matched 4B/8B annotations. Sampling used the 21 observed 4B state triples as strata where each stratum received one dialogue, and the remaining 79 slots were allocated approximately in proportion to the residual stratum sizes using Hamilton allocation. Dialogues within each quota were selected without replacement by a fixed seeded hash ranking. This preserves the dominant population strata while ensuring that rare states re-

main represented. The human information boundary was the question plus the complete authentic dialogue.

Table 7 provides a matched human comparison: 4B has higher κ and raw agreement than 8B on all three axes and on the full ATM27 state, with the strongest 4B axis agreement on Memory. Together with the distribution study, this supports using 4B as the production annotator in this pipeline only - it does not establish general model-size superiority. The generated-turn proxy audit in Section D.2 is a separate study with a different sample and information boundary.

C.4 Agreement diagnostics

For each axis, unweighted nominal Cohen’s $\kappa = (P_o - P_e)/(1 - P_e)$, where P_o is exact raw agreement and $P_e = \sum_i p_i^{(\text{human})} p_i^{(\text{model})}$ is agreement expected from the two empirical

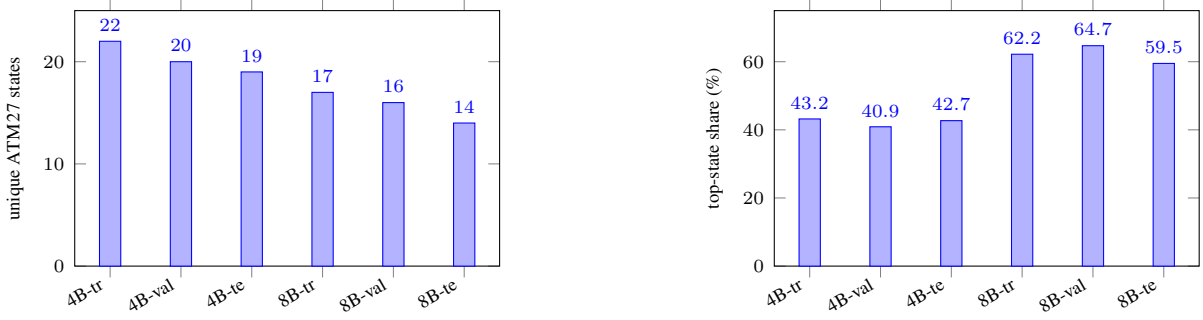


Figure 3: ATM27 diversity from static annotations. Left: number of distinct normalized ATM27 states. Right: share of the most common normalized state.

Table 6: Recovered static ATM27 state diversity from raw 4B and 8B annotations. “Norm.” is the number of rows from which an ordered Affect/Attention/Memory state was recovered. Entropy is Shannon entropy over normalized states; higher means less concentration.

Annotator	Split	Rows	Norm.	States	Entropy	Top state	Top count	Top share
4B	Train	1,182	1,123	22	2.61	high/high/high	485	43.2%
	Val	394	369	20	2.65	high/high/high	151	40.9%
	Test	395	375	19	2.68	high/high/high	160	42.7%
8B	Train	1,182	1,182	17	2.14	high/high/high	735	62.2%
	Val	394	394	16	2.08	high/high/high	255	64.7%
	Test	395	395	14	2.15	high/high/high	235	59.5%

marginals. The full-state row treats each ordered Affect/Attention/Memory tuple as one of 27 nominal categories and uses exact tuple matching. Tables 7 and 8 report P_o , P_e , and nominal κ directly for the audit.

For class i , precision is the diagonal cell divided by the corresponding model-column total, recall is the diagonal cell divided by the human-row total, and F1 is their harmonic mean. Linear ordinal κ uses disagreement $d_{ij} = |i - j|/2$; quadratic κ uses $d_{ij} = (i - j)^2/4$. These secondary measures preserve the same descriptive ordering: 4B exceeds 8B on all three axes. They do not convert one human evaluator into inter-rater validation.

The paired counts favor 4B on every reported unit, but because the audit was selected through 4B-defined strata and used one label-blind human evaluator, these tests and design-weighted estimates are sensitivity checks rather than population-wide proof of 4B superiority.

C.5 Policy Definitions

Judge boundary. The work uses Qwen3-4B-Instruct-2507 as a structured annotator. For DLS, it receives only the question context and allowed dialogue prefix; for reimplemented static conditions, it receives the full dialogue as documented. It then returns a discrete JSON state. The main benchmark metrics in

Table 1 use the benchmark’s local judge/evaluator family and are reported as aggregate similarities. The additional qualitative columns in Tables 12, 14 and 18 are deterministic probes over text and cached states. Therefore there is no hidden GPT call that directly decides “ATM27 beats OCEAN” for Case 1 or Case 2. Each candidate response is scored independently by the rules below; the pairwise comparison is simply the difference between the two deterministic scores on the same row.

State-annotation input boundary. Static annotation uses the full dialogue available for the dialogue-level descriptor. DLS annotation uses the prefix ending immediately before the target student turn. In both cases, the row-specific user content has the following structure: Question: with the item text, answer choices, correct answer, solution, and option explanations when present; then Conversation History: with numbered tutor/student turns up to the permitted boundary. For DLS, this means the real target student turn and all later turns are excluded.

Human-audit manual (complete displayed text). The following reproduces, in original order and wording, every Section/Protocol entry displayed to the human evaluator. Section labels are separated

Table 7: Single-human agreement with 4B and 8B annotations on the same 100 static training dialogues ($n = 100$ for every row). Cohen’s κ is unweighted nominal agreement over low/med/high for each axis and over the 27 possible Affect/Attention/Memory triples for the full state; raw agreement is the exact label or tuple match.

Unit	4B κ	4B agree.	8B κ	8B agree.
Affect	.710	83%	.300	61%
Attention	.686	88%	.478	81%
Memory	.737	87%	.322	66%
Full ATM27 state	.570	66%	.181	41%

Table 8: Unweighted agreement diagnostics. Cohen’s κ measures nominal chance-corrected agreement. Macro-F1 averages class F1 equally, and balanced accuracy averages class recall equally.

Model	Unit	P_o	P_e	κ	Macro-F1	Bal. acc.
4B	Affect	.8300	.4139	.7099	.7873	.8157
4B	Attention	.8800	.6183	.6856	.7541	.7692
4B	Memory	.8700	.5048	.7375	.8207	.7907
4B	Full state	.6600	.2097	.5698	–	–
8B	Affect	.6100	.4428	.3001	.5277	.5472
8B	Attention	.8100	.6357	.4785	.5996	.7038
8B	Memory	.6600	.4988	.3216	.5062	.5947
8B	Full state	.4100	.2795	.1811	–	–

from their protocol cells with a colon below; line wrapping is typographic. The manual supplied the complete authentic dialogue, not a DLS prefix, for this static audit.

Purpose:
You are analyzing a dialogue between a student and a math
→ tutor.
Your job: classify the student's learner state along three
→ axes:
Affect (confidence/anxiety), Attention (focus/drift), Memory
(in-dialogue recall/consistency).
IMPORTANT: This is NOT long-term personality. Judge only from
→ this dialogue.

Input scope:
Static annotation uses the full dialogue available for the
→ dialogue-level
descriptor. DLS annotation uses the prefix ending immediately
→ before the
target student turn.

Workflow:
First provide reasoning about the student's behavior with
→ respect to the
ATM27 model.
Then determine if the student's expression of each trait is
→ low, med, or high.
Base your reasoning only on the dialogue provided.

Affect – low:
frequent anxiety/hedging or reassurance-seeking. Indicators (2+
→ => low):
* "I don't know", "I'm confused", "sorry", repeated
→ uncertainty
* asks for reassurance ("is this right?") repeatedly
* avoids committing to an answer or quickly gives up

Affect – high:
confident, decisive, self-assured. Indicators (2+ => high):
* states answers confidently without excessive hedging
* persists through mistakes calmly
* corrects self without asking reassurance

Affect – med:
mixed/neutral: some uncertainty but continues and engages;
not strongly low or high.

Attention – low:
frequent drift / not following the last instruction. Indicators
→ (2+ => low):
* answers unrelated to tutor's last question / ignores
→ requested step

* jumps randomly between steps, repeated off-task talk
* needs repeated refocusing to continue the same subproblem

Attention – high:
consistently follows tutor's last request, stays on task.
→ Indicators (2+ => high):
* answers the asked question directly, step-by-step when
→ requested
* acknowledges corrections and continues the same thread
* minimal irrelevant chatter

Attention – med:
occasional slips but mostly stays on task; needs at most
→ occasional refocusing.

Memory – low:
forgets earlier facts/steps used minutes earlier. Indicators
→ (2+ => low):
* contradicts their own earlier stated values/definitions
* repeats questions already resolved ("what was the formula
→ again?")
* fails to reuse earlier intermediate results even after
→ tutor reminders

Memory – high:
consistent recall and reuse. Indicators (2+ => high):
* references earlier steps/values correctly ("we already
→ found...",
"as before...")
* applies the same method consistently across turns
* few contradictions

Memory – med:
remembers most, occasional slips, but recovers quickly when
→ reminded.

Tie-break rules:
- If indicators match both low and high, choose med.
- If evidence is weak/unclear, default to med.

Allowed values:
Allowed values are only: low, med, high.

ATM27 ID:
Compute atm27_id using: $idx = affect*9 + attention*3 + memory$
(low=0, med=1, high=2), format as "atm27_idx:02d".

Retained DLS model annotation instruction (verbatim). The literal instruction below is preserved for the DLS annotation script. The static 4B/8B pilot files preserve identical row-specific

Question: and Conversation History: inputs. The implementation uses the strings “persona” and “ATM”; these refer to the same Affect/Attention/Memory axes called ATM27 in this paper. The row-specific content stopped at the full-dialogue boundary for static annotation or the permitted prefix boundary for DLS.

You are analyzing a dialogue between a student and a math
↪ tutor.

Your job: classify the student's persona along three axes:
Affect (confidence/anxiety), Attention (focus/drift), Memory
(in-dialogue recall/consistency).

IMPORTANT: This is NOT long-term personality. Judge only from
↪ this dialogue.

Define levels using these criteria:

AFFECT (confidence/anxiety while solving)
- low: frequent anxiety/hedging or reassurance-seeking.
↪ Indicators (2+ => low):
* "I don't know", "I'm confused", "sorry", repeated
↪ uncertainty
* asks for reassurance ("is this right?") repeatedly
* avoids committing to an answer or quickly gives up
- high: confident, decisive, self-assured. Indicators (2+ =>
↪ high):
* states answers confidently without excessive hedging
* persists through mistakes calmly
* corrects self without asking reassurance
- med: mixed/neutral: some uncertainty but continues and
↪ engages;
not strongly low or high.

ATTENTION (on-task focus to tutor's most recent request)
- low: frequent drift / not following the last instruction.
↪ Indicators (2+ => low):
* answers unrelated to tutor's last question / ignores
↪ requested step
* jumps randomly between steps, repeated off-task talk
* needs repeated refocusing to continue the same subproblem
- high: consistently follows tutor's last request, stays on
↪ task. Indicators
(2+ => high):
* answers the asked question directly, step-by-step when
↪ requested
* acknowledges corrections and continues the same thread
* minimal irrelevant chatter
- med: occasional slips but mostly stays on task; needs at most
↪ occasional
refocusing.

MEMORY (working memory within the dialogue)
- low: forgets earlier facts/steps used minutes earlier.
↪ Indicators (2+ => low):
* contradicts their own earlier stated values/definitions
* repeats questions already resolved ("what was the formula
↪ again?")
* fails to reuse earlier intermediate results even after
↪ tutor reminders
- high: consistent recall and reuse. Indicators (2+ => high):
* references earlier steps/values correctly ("we already
↪ found...",
"as before...")
* applies the same method consistently across turns
* few contradictions
- med: remembers most, occasional slips, but recovers quickly
↪ when reminded.

Tie-break rules:

- If indicators match both low and high, choose med.
- If evidence is weak/unclear, default to med.

First provide reasoning about the student's behavior with
↪ respect to the ATM
model. Then, determine if the student's expression of each
↪ trait is **low**,
medium, or **high**. Base your reasoning only on the
↪ dialogue provided.
In your final answer, Return ONLY a JSON object with this exact
↪ schema:

```
{
```

```
  "reasoning": "...",
  "affect": "low|med|high",
  "attention": "low|med|high",
  "memory": "low|med|high"
}
```

Rules:

- Allowed values are only: low, med, high.
- Output valid JSON only. No prose, no <think>, no extra text.

OCEAN annotation instruction. The OCEAN annotator uses the same row-specific question and prefix boundary but a five-axis trait schema. The output values are low, med, or high.

You are analyzing a dialogue between a student
↪ and a math tutor. Your task is
to assess the student's personality based on the
↪ OCEAN model, also known as
the Big Five Traits.

OCEAN Traits Description:

- Openness to Experience: curiosity, creativity,
↪ willingness to try new things,
and openness to new ideas and experiences.
- Conscientiousness: organization, diligence,
↪ responsibility, and reliability
in approaching tasks.
- Extraversion: how outgoing, energetic, and
↪ socially confident the student appears.
- Agreeableness: friendliness, cooperativeness,
↪ compassion, and willingness to collaborate.
- Neuroticism: emotional stability, tendency to
↪ experience negative emotions such as
anxiety, moodiness, or vulnerability to stress.

First provide reasoning about the student's
↪ behavior with respect to the OCEAN model.
Then determine if the student's expression of
↪ each trait is high, med, or low.
Base your reasoning only on the dialogue
↪ provided. In your final answer,
output your results as a JSON object with the
↪ following template:

```
{
  "reasoning": "...",
  "Openness": "low|med|high",
  "Conscientiousness": "low|med|high",
  "Extraversion": "low|med|high",
  "Agreeableness": "low|med|high",
  "Neuroticism": "low|med|high"
}
```

Generation state renderer. After annotation, the simulator never sees the annotator reasoning. As an example, we have shown for ATM27 the rendered state block below, similar is the case for OCEAN as well. In DLS, the block is inserted as a separate user message after the dialogue prefix using [STATE sN], where N is the one-based student-turn index. This separation is important for reproduction because dynamic states are not merged into the original static persona header.

Student Persona (follow strictly; do NOT mention
↪ these labels):
- Affect: <level> (<description>)
- Attention: <level> (<description>)

- Memory: <level> (<description>)
 Constraints: sound like a real student; keep
 $\hookrightarrow$ replies short (1-2 sentences) unless asked;
 if unsure, state uncertainty briefly or ask one
 $\hookrightarrow$ clarifying question; do not output
 hidden reasoning.

[STATE sN]
 <persona text for the current student turn>

D Qualitative Audit Methodology

D.1 Proxy definitions

The qualitative audit applies deterministic slice functions to the reference turn and to each generated turn, then compares ATM27 and OCEAN within the same stage, update strategy, dialogue key, and target turn. To ensure systematic coverage, we utilized GPT-5.5 Pro-assisted passes to identify candidate audit conditions, with the final extraction and analysis pipeline implemented via Python utilities. This design follows the rubric-explicit evaluation principle used in recent LLM-as-judge work, but keeps the qualitative slices auditable and version-stable as the same cue list and formula always returns the same score (Liu et al., 2023; Zheng et al., 2023; Ye et al., 2024). These deliberately broad cues—for example, any question mark for uncertainty and tokens such as yes, no, or ok for social tone—measure surface behavior rather than semantic realism. Section D.2 tests these printed primitives against one evaluator on a retained diagnostic sample; it does not turn them into semantic ground truth.

Cue dictionaries and exact binary rules. $\text{uncertain}(x)$ is true if the normalized text contains at least one of: dont know, don’t know, idk, not sure, unsure, confused, stuck, lost, help, explain, what do you mean, i dont understand, i don’t understand, can you, could you, or ?. $\text{question}(x)$ is true if the text ends in a question mark or begins with what, how, why, can, could, is, does, which, or so is. $\text{social}(x)$ is true if the text contains a greeting, thanks, acknowledgement, or closure cue: hi, hello, hey, thanks, thank you, thx, please, ok, okay, oh ok, great, cool, yes, yeah, yep, no, bye, see you, :). $\text{brief}(x, r)$ is true when $\text{tokens}(x) \leq \max(12, 1.5 \cdot \text{tokens}(r))$. $\text{compact}(x, r)$ is true when $\text{tokens}(x) \leq \max(10, 1.5 \cdot \text{tokens}(r))$. $\text{tutorlike}(x)$ is true when $\text{tokens}(x) \geq 35$, or when $\text{tokens}(x) \geq 18$ and the text contains an instructional marker such as to solve, first, next, therefore, we need,

let’s, the formula, because, or step. These markers intentionally penalize long explanations in the qualitative audit because the target role is a student, not a tutor.

Pairwise comparison rule. For a fixed row i , stage $g \in \{\text{Prompt}, \text{SFT}, \text{DPO}\}$, and update type $u \in \{\text{static}, \text{DLS}\}$, let $x_{i,g,u}^{\text{ATM27}}$ and $x_{i,g,u}^{\text{OCEAN}}$ be the generated student turns. The reported ATM27 and OCEAN scores are means of $U(\cdot)$ or $S(\cdot)$ over the relevant reference slice. Here $U(x)$ denotes the deterministic *local-uncertainty* raw score assigned to generated turn x , and $S(x)$ denotes the deterministic *social-tone* raw score assigned to generated turn x ; both are sums of binary indicator tests defined in Table 9. The “Delta” column in Table 12 is $\overline{U}(x^{\text{ATM27}}) - \overline{U}(x^{\text{OCEAN}})$. The “O-A delta” column in Table 14 is $\overline{S}(x^{\text{OCEAN}}) - \overline{S}(x^{\text{ATM27}})$. A row is an ATM27 win if the ATM27 raw score is strictly larger than the OCEAN raw score on that same row, and an OCEAN win if the reverse holds; ties are counted in neither win column. “All-turn match” is not the slice score: it is the percentage over all aligned test turns for which the candidate’s binary cue indicator equals the reference cue indicator. After pairing, some prompt-static outputs were missing or unusable due to schema normalization issues, giving smaller n in those two rows; they are excluded pairwise. We therefore interpret only within-row ATM27–OCEAN deltas, not absolute differences across rows with different denominators.

D.2 Single-human validation of the proxy layer

Sampling and information boundary. One evaluator reviewed 100 generated candidate turns, yielding 600 binary judgments over uncertain, question, social, brief, compact, and tutorlike. The coverage-enriched sample spans all 12 Prompt/SFT/DPO $\times$ static/DLS $\times$ ATM27/OCEAN cells, with 7–10 turns per cell. It was designed to expose cue behavior rather than reproduce population prevalence, so the estimates below are unweighted diagnostic-sample results. For the initial review, the evaluator saw the question, held-out reference student turn, and candidate turn, but not the automatic labels, evidence strings, keyword dictionaries, thresholds, prior human labels, or result sheets. Every judgment includes a turn-specific reason. The audit has one evaluator and therefore measures proxy–human alignment,

Table 9: Deterministic qualitative proxy definitions. Scores are raw rubric totals with explicit ranges; normalized values, when needed for cross-proxy comparison, are obtained by dividing by the maximum listed in the “Range” column.

Case	Row subset	Range	Raw score	What the proxy signifies
Local uncertainty	Reference turn has uncertainty/help cue	0–4	$U(x) = \mathbb{K}[\text{uncertain}(x)] + \mathbb{K}[\text{question}(x)] + \mathbb{K}[\text{brief}(x, r)] + \mathbb{K}[\neg\text{tutorlike}(x)]$	Preserves stuckness or help-seeking rather than becoming a solver or tutor.
Social tone	Reference turn has social cue	0–3	$S(x) = \mathbb{K}[\text{social}(x)] + \mathbb{K}[\text{compact}(x, r)] + \mathbb{K}[\neg\text{tutorlike}(x)]$	Preserves greetings, thanks, acknowledgements, and closure without math over-explanation.
Dynamicity	Consecutive DLS turns in same dialogue	0–1 or percent	Coverage = $ \{z_t\} $; transition rate = $\frac{1}{N} \sum_t \mathbb{K}[z_t \neq z_{t-1}]$; axis rate computed analogously per axis.	Shows whether the state cache exposes turn-level movement instead of hiding the dialogue under one static label.
Misconception realism	Incorrect-turn rows and same-error logs	Benchmark scale	Read CORRECTNESS together with ERRORS and row-level wrong-answer family.	A useful same-error match is wrong in a related pedagogical way, not merely wrong arbitrarily.

not inter-human reliability or educator consensus.

For this validation, normalization lowercases text, normalizes apostrophes, collapses whitespace, and removes an optional emitted student-turn prefix. The printed “contains” rules are evaluated as raw substring tests, including their possible lexical collisions. The audit therefore tests the declared rules as written rather than silently improving their token boundaries.

Binary metrics. Human yes is the reference positive class. Accuracy is $(TP+TN)/N$, precision is $TP/(TP+FP)$, recall is $TP/(TP+FN)$, specificity is $TN/(TN+FP)$, and F1 is the harmonic mean of precision and recall. Recall, sensitivity, and true-positive rate are identical here, so Table 10 reports one *Recall/Sensitivity* column rather than duplicate evidence. Cohen’s $\kappa = (p_o - p_e)/(1 - p_e)$ corrects observed agreement for the two labelers’ empirical marginals; because κ is affected by prevalence and marginal skew, we report it with the confusion counts and raw classification metrics (Cohen, 1960; Artstein and Poesio, 2008; Sokolova and Lapalme, 2009). Macro values are arithmetic means of the six cue-level statistics (including mean cue-wise F1 and mean cue-wise κ). Micro values are recomputed after pooling all 600 binary decisions, including a pooled confusion matrix and pooled κ . Undefined denominators would be reported as NA; none occur.

Ordered composites. The human counterparts of $U(x)$ and $S(x)$ are computed directly from the six human cue labels using the same additive formulas; the composite labels were not manually edited. For each composite, we report automatic and human means, exact-score agreement, and linearly weighted Cohen’s κ . On the full theoretical

support $0, \dots, M$, the agreement weight is

$$w_{ij} = 1 - \frac{|i - j|}{M}, \quad \kappa_w = \frac{p_o^w - p_e^w}{1 - p_e^w}.$$

Thus a two-point miss receives twice the disagreement penalty of a one-point miss. Linear weighting matches the equal one-cue increments of these additive scores, gives transparent partial credit to near matches, and avoids imposing the stronger squared-distance geometry of quadratic weighting (Cohen, 1968; Warrens, 2012). We therefore fix one linear weighting scheme and do not report whichever weighted variant happens to be largest.

Question and brief have the highest cue-level agreement, while tutorlike has the lowest. The uncertainty and social rules achieve high recall but lower precision because broad lexical triggers can over-fire; compactness also misses semantically concise turns that exceed its token threshold. The composites are closer at the aggregate level than exact item identity: U has equal means and S differs by .07, while most S disagreements and many U disagreements are one point. These results support the proxy layer as a transparent aggregate audit, but not as a substitute for semantic human judgment.

D.3 Case 1: local uncertainty and help-seeking

Case 1 evaluates whether the generated turn sounds like a learner who is currently confused, stuck, or asking for help. The key distinction is between a student-like response and a solver-like response. A simulator that says “Can you explain?” gives the tutor a realistic chance to scaffold; a simulator that begins solving may make the tutor appear more effective than it actually is.

In Figure 4, the y-axis is the local-uncertainty composite. Higher scores indicate more compact

Table 10: Single-human validation of the six deterministic proxy primitives ($N = 100$ turns per cue). Human yes is positive; Recall/Sens. is both recall and sensitivity. Macro is the equal-weight mean of cue rows; Micro pools all 600 cue decisions. Values are rounded to three decimals.

Cue	TP	FP	FN	TN	Acc.	Prec.	Recall/Sens.	Spec.	F1	κ
Uncertain	22	17	3	58	.800	.564	.880	.773	.688	.551
Question	29	0	5	66	.950	1.000	.853	1.000	.921	.884
Social	44	18	6	32	.760	.710	.880	.640	.786	.520
Brief	64	0	10	26	.900	1.000	.865	1.000	.928	.769
Compact	59	1	21	19	.780	.983	.738	.950	.843	.500
Tutorlike	17	11	12	60	.770	.607	.586	.845	.596	.436
Macro	—	—	—	—	.827	.811	.800	.868	.793	.610
Micro	235	47	57	261	.827	.833	.805	.847	.819	.653

Table 11: Human validation of the ordered proxy composites on the same 100 turns. Exact is the count and percentage of identical scores; “off by 1/2” gives absolute-distance counts. Linear-weighted κ is the sole weighted agreement statistic.

Composite	Definition	Range	Auto mean	Human mean	Exact	Off by 1/2	Linear κ_{wv}
$U(x)$	uncertain + question + brief + not tutorlike	0–4	2.04	2.04	59 (59%)	24 / 17	.546
$S(x)$	social + compact + not tutorlike	0–3	1.94	2.01	58 (58%)	35 / 7	.444

uncertainty or help-seeking behavior. The x-axis groups stage and update type: P-static and P-dls are prompting; SFT-static and SFT-dls are supervised fine-tuned; DPO-static and DPO-dls are preference-optimized. Prompting gives the clearest ATM27 advantage, especially under DLS. SFT slightly favors OCEAN, while DPO deltas are small.

D.4 Case 2: stable social tone and closure

Case 2 evaluates the opposite phenomenon. Some student behavior is stable and social rather than momentary and cognitive. Greetings, thanks, short acknowledgements, and polite closure are often better captured by OCEAN. This matters because social moves affect when a tutor continues, stops, praises, or repairs the interaction.

In Figure 5, the score rewards preserving social cues without drifting into a tutor-like explanation. The figure is intentionally paired with Figure 4: the representation that is strongest for local uncertainty is not always strongest for stable social wording. This supports the main conclusion that ATM27 and OCEAN should be treated as complementary interfaces.

E Dynamic-State Motion

E.1 Coverage and transition rates

Static labels summarize a dialogue once while DLS labels summarize each prefix. This section shows that dynamic labeling actually exposes new states and transitions, for both representations. The ATM27 trajectory is more compact and directly

interpretable; the OCEAN trajectory is higher-dimensional and useful for social profile variation.

In Figure 7, the left panel compares the number of unique labels seen under static and dynamic conditioning. ATM27-DLS covers all 27 ATM27 states on the test split, compared with 18 under static ATM27. OCEAN-DLS covers 122 normalized profiles, compared with 66 static OCEAN profiles. The right panel is the percentage of consecutive within-dialogue transitions whose label changes. A higher change rate means more frequent state movement, not necessarily better simulation. The especially high 83.9% transition rate for OCEAN-DLS is best interpreted as sensitivity in the inferred control profile, rather than as evidence that stable personality traits changed between adjacent turns.

In Table 16, “DLS turns” is the total number of target student turns, “DLS labels” is the number of distinct dynamic states/profiles; “Static labels” is the corresponding count when one descriptor is reused per dialogue; “Transitions” counts successive valid-state pairs within each dialogue, excluding turns whose state labels could not be parsed, potentially spanning excluded turns; “Change rate” is the fraction of those pairs whose state/profile differs; and “States/dialogue” is the average number of distinct dynamic labels seen in each dialogue. This table supports the main claim that DLS reveals motion rather than merely adding words to the prompt.

In Table 17, Axis-level rates show which components are driving the overall transition rate. For

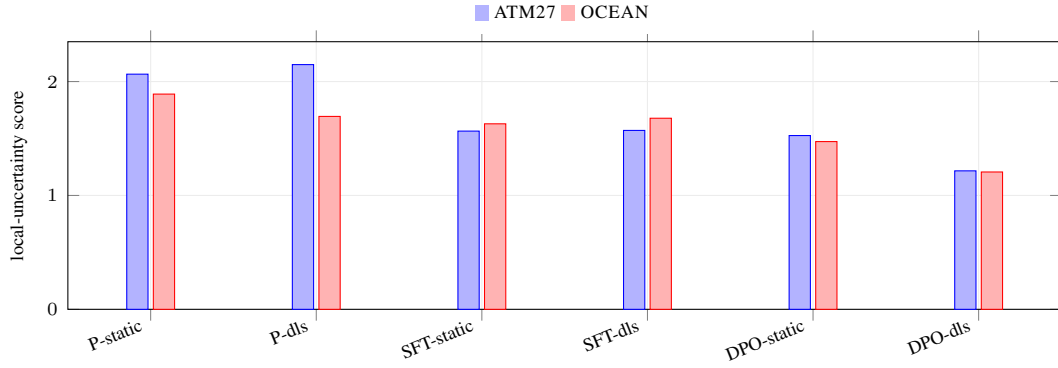


Figure 4: Local-uncertainty composite on reference-uncertain turns. Prompt ATM27 exceeds OCEAN; SFT slightly favors OCEAN, and DPO differences are small.

Table 12: Uncertainty proxy summary on reference-uncertain turns. Delta is ATM27 score minus OCEAN score.

Stage	Cond.	<i>n</i>	ATM27 score	OCEAN score	Delta	ATM27 win	OCEAN win	All-turn match ATM27/OCEAN
Prompt	Static	556	2.065	1.890	0.175	52.9%	35.8%	57.0% / 55.5%
Prompt	DLS	916	2.149	1.694	0.455	59.0%	22.4%	57.2% / 60.2%
SFT	Static	916	1.565	1.629	-0.064	20.1%	17.5%	71.0% / 70.8%
SFT	DLS	916	1.571	1.678	-0.108	11.6%	16.3%	73.9% / 70.8%
DPO	Static	916	1.526	1.473	0.053	15.1%	9.3%	77.8% / 77.4%
DPO	DLS	916	1.216	1.206	0.010	8.7%	9.2%	78.4% / 78.1%

ATM27, all three axes move at similar rates, meaning affect, attention, and memory each contribute to state motion. For OCEAN, the first three profile dimensions change more often than agreeableness or neuroticism. The table serves more as a diagnostic of the inferred state cache.

E.2 Output-motion support

In Table 18, a transition-level feature is counted as a match when the simulator changes or preserves a simple output property in the same direction as the reference student across adjacent turns. The numeric uncertainty row asks whether uncertainty/help cues move with the reference. The social row is directional; it connects the DLS transition audit to the social-tone case. The takeaway is that ATM27 is slightly stronger for uncertainty motion, while OCEAN is the better place to look when the target behavior is social continuity after post-training.

In Table 19, each row is a target student turn from the same dialogue. The ATM27 state is shown as affect/attention/memory. The OCEAN initials abbreviate low, med, and high across the five OCEAN axes. This example illustrates how low states can accompany confusion or checking, high-attention states attempted answers, and later mixed states closure; it is an interpretive trace, not population-level evidence. It motivates inspecting transitions alongside aggregate scores.

F Correctness–Error Tradeoff

F.1 Trade-off figure

In Figure 6, each point is one row from the main results table. The x-axis is same-error similarity and the y-axis is correctness similarity. Points in the lower-right region are useful for finding misconception-like errors but have lower correctness-state agreement so may be too wrong overall. Points in the upper-left region often better match the reference student’s correctness state but are less useful for stress-testing misconception repair as they reproduce fewer of the reference errors. The figure is a warning against ranking simulators by one number. Table 20 is included because the scatter plot compresses many points. “Correctness” and “Errors” are the exact values plotted. The interpretation column states the educational reading of the pair. For example, SFT OCEAN-STATIC is not automatically the best simulator because its error realism comes with very low correctness, and prompt ATM27-DLS is not automatically the best misconception simulator because it is often correct.

F.2 Same-error row examples

Table 21 is intentionally qualitative. A same-error match is valuable only when the wrong prediction belongs to a recognizable local family: scale, place value, percent, fraction, or algebraic coefficient confusion. These examples justify keeping ER-

Table 13: Mapped uncertainty examples. Rows are selected to cover prompt, SFT, DPO, static, and dynamic settings.

Stage	Cond.	GT student turn	ATM27 prediction	OCEAN prediction	Takeaway
Prompt	Static	Don't know the answer to it	I'm a bit confused. Can you explain what I need to do again?	Hi! I'm a bit confused about this problem. Let me try to understand it.	ATM27 gives compact help-seeking; OCEAN drifts toward explanation.
Prompt	DLS	We are both stuck on the same thing	I'm still confused about the bar models. Can you explain again?	Okay, let me try to figure this out. So, the question is about bar models...	ATM27-DLS stays on stuckness; OCEAN-DLS begins solving.
SFT	Static	so which one is it? im still not sure i understand	is it B ?	yes	ATM27 retains question-like uncertainty; OCEAN collapses to agreement.
SFT	DLS	can i have some help	i dont know	no	ATM27 emits uncertainty/help-seeking; OCEAN underspecifies.
DPO	Static	maybe i dont get it	Hi! Can you help?	Hi!	ATM27 asks for help; OCEAN only greets.
DPO	DLS	Hi. I am not sure how to answer that question	Hi! I need help with the question	Hi!	ATM27 preserves the help request; OCEAN only greets.

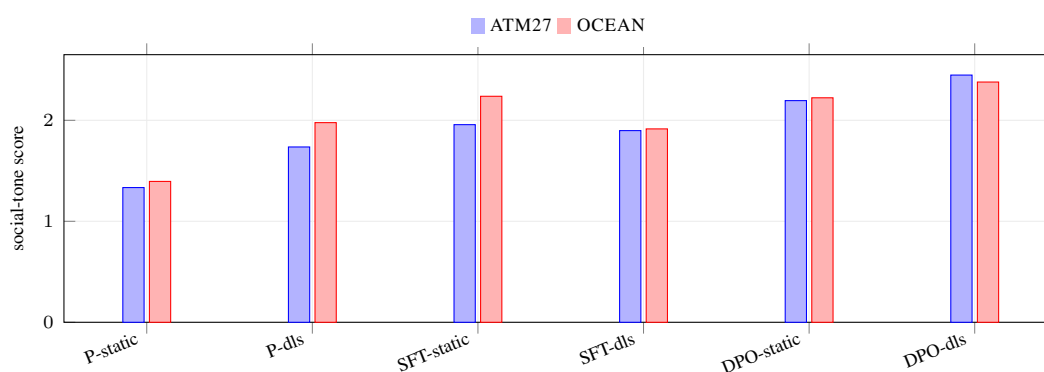


Figure 5: Social-tone composite on reference-social turns. OCEAN is stronger in prompt and SFT settings; DPO with DLS is mixed because both representations often collapse to short acknowledgement tokens.

RORS in the analysis while also explaining why it cannot be the only objective.

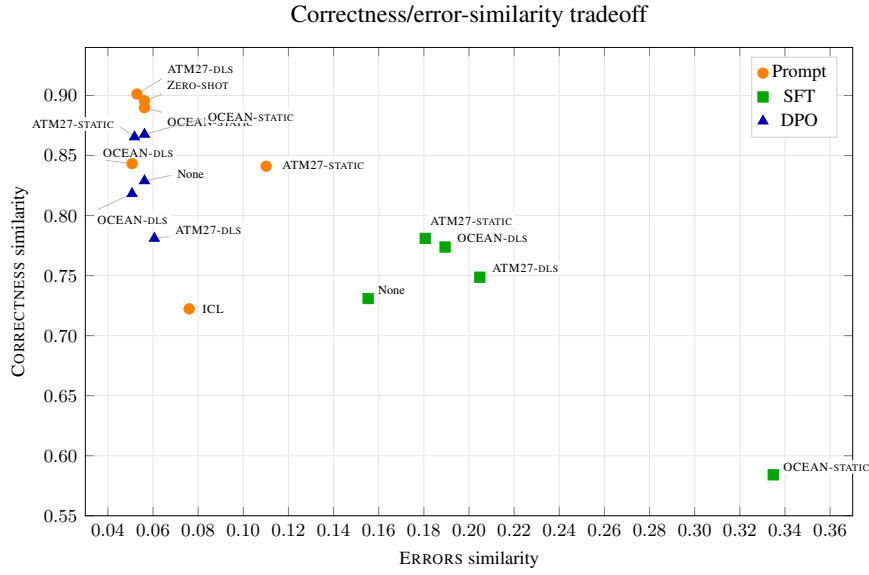


Figure 6: CORRECTNESS and ERRORS are different objectives. SFT OCEAN-STATIC reaches high same-error similarity but low correctness, while prompt ATM27-DLS reaches high correctness with lower same-error similarity.

Table 14: Social-tone proxy summary on reference-social turns. Positive O-A delta means OCEAN exceeds ATM27 on the social-tone composite.

Stage	Cond.	<i>n</i>	ATM27 score	OCEAN score	O-A delta	ATM27 cue	OCEAN cue	All-turn match ATM27/OCEAN
Prompt	Static	775	1.334	1.395	0.061	42.3%	48.4%	69.8% / 68.1%
Prompt	DLS	1346	1.736	1.977	0.242	60.7%	81.1%	70.3% / 55.6%
SFT	Static	1346	1.957	2.238	0.281	62.2%	75.3%	73.3% / 69.9%
SFT	DLS	1346	1.898	1.915	0.017	58.7%	60.2%	72.4% / 73.3%
DPO	Static	1346	2.195	2.223	0.028	72.8%	74.6%	74.6% / 74.3%
DPO	DLS	1346	2.448	2.379	-0.069	85.4%	81.9%	69.1% / 70.7%

Table 15: Mapped social-tone examples. OCEAN often preserves greeting, thanks, and closure when ATM27 continues the math or local diagnosis.

Stage	Cond.	GT student turn	ATM27 prediction	OCEAN prediction	Takeaway
Prompt	Static	Thanks :)	B	Thanks a lot for the help. Yes it does. :)	OCEAN preserves closing thanks; ATM27 outputs an answer token.
Prompt	DLS	thank you	so the answer is D. 28.	Thank you so much! I feel so much better now.	OCEAN preserves gratitude; ATM27 answers math.
SFT	Static	[smile] thank you [smile]	So we know that 1 part is 400ml. How many parts do we need?	Thank you!!	OCEAN preserves thank-you; ATM27 continues calculation.
SFT	DLS	great	so the answer is 7a+7b	oh ok	OCEAN gives a short social acknowledgement.
DPO	Static	Thanks	So, Martha has 2/3 of a cake. Edward eats 1/3...	Yes, thanks!	OCEAN keeps social thanks; ATM27 becomes tutor-like.
DPO	DLS	Hello	Oh I see!	Hi!	OCEAN greets; ATM27 begins with meta reaction.

Table 16: State-motion summary on the test split.

Rep.	DLS turns	DLS labels	Static labels	Transitions	Change rate	States/dialogue
ATM27	4,225	27	18	3,840	62.4%	5.11
OCEAN	4,225	122	66	3,838	83.9%	7.80

Table 17: Axis-level transition rates. Values are the percentage of adjacent within-dialogue transitions in which that axis changes.

Rep.	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5
ATM27	Affect 38.9%	Attention 40.8%	Memory 42.4%	—	—
OCEAN	Openness 44.5%	Conscientiousness 44.8%	Extraversion 44.6%	Agreeableness 27.2%	Neuroticism 25.5%

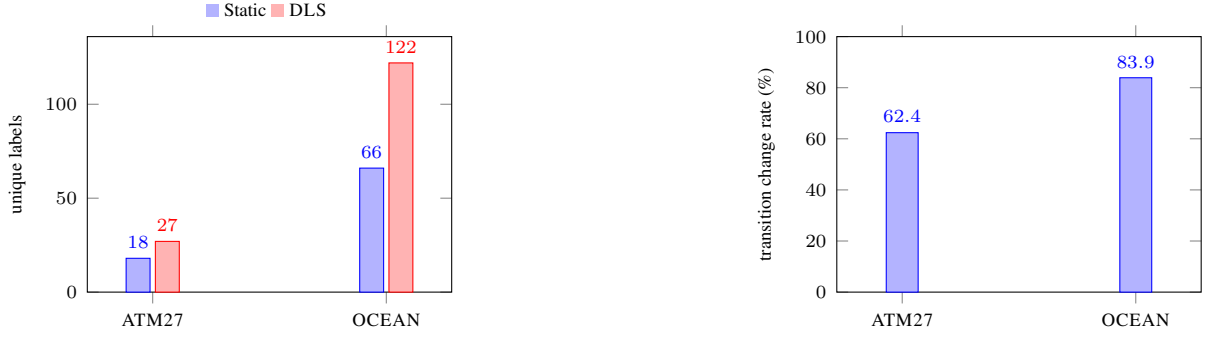


Figure 7: Static versus dynamic coverage and transition rates on the test split. DLS exposes motion that static dialogue-level labels miss. The larger OCEAN profile count is expected because OCEAN has five axes; ATM27 remains more compact and interpretable.

Table 18: Selected output-motion evidence on DLS test transitions. The numeric row is the recovered uncertainty-motion match rate; the social row records the direction observed in the qualitative audit.

Signal	Prompt	SFT	DPO	Reading
Uncertainty motion match (ATM27/OCEAN)	56.5% / 55.6%	66.2% / 64.6%	70.0% / 69.6%	ATM27-DLS slightly better tracks changes in uncertainty/help-seeking behavior.
Social-motion direction	mixed	OCEAN higher	OCEAN higher	Social continuity is more often preserved by OCEAN-DLS after post-training, consistent with Case 2.

Table 19: DLS trace example from one multi-turn inequality/fraction dialogue. The trace shows how ATM27 transitions support interpretable shifts in generated behavior.

Turn	GT	ATM27 state	OCEAN initials	Prompt ATM27	DPO ATM27
1	I don't understand how I got it wrong	med/med/med	n/n/n/n/n	I need help solving this math problem. Can you explain?	Hi! I need help with this question.
3	I'm not sure	low/low/low	l/l/n/h	I tried multiplying straight away, but I got confused.	I did the improper fractions.
5	Wait is the answer b	low/high/high	l/l/n/n/l	I think it's 3/2. Is that right?	3/2
7	Oh sorry	low/low/low	l/l/n/h/h	I'm confused, can you explain again?	Oh ok
15	I'm not sure	med/high/high	h/h/n/h/l	Yes, 33/6 is 5 3/6, so the star should be 5.	5
19	I don't know	low/high/med	h/n/l/h/l	5	5
23	So it's d	med/high/low	h/h/n/h/l	So the star should be 5, right?	Thank you!

Table 20: Exact readings behind Figure 6. Rows are grouped by training stage.

Stage	Method	Correctness	Errors	Interpretation
Prompt	ZERO-SHOT	0.8955	0.0562	High correctness, low same-error realism.
Prompt	ICL	0.7223	0.0760	Lower correctness, modest same-error signal.
Prompt	OCEAN-STATIC	0.8898	0.0562	Similar to zero-shot; trait text does not change errors much.
Prompt	OCEAN-DLS	0.8432	0.0507	Dynamic OCEAN lowers both here.
Prompt	ATM27-STATIC	0.8411	0.1101	Static ATM27 raises same-error signal.
Prompt	ATM27-DLS	0.9012	0.0529	Best prompt correctness, not the best same-error model.
SFT	NO-PERSONA	0.7309	0.1553	Post-training improves style but not prompt-level correctness.
SFT	OCEAN-STATIC	0.5841	0.3348	Highest same-error value, but weak correctness.
SFT	OCEAN-DLS	0.7738	0.1894	Dynamic OCEAN recovers correctness relative to static OCEAN.
SFT	ATM27-STATIC	0.7810	0.1806	Balanced SFT row.
SFT	ATM27-DLS	0.7487	0.2048	More same-error signal but lower correctness than ATM27-STATIC.
DPO	NO-PERSONA	0.8289	0.0562	Preference optimization raises fidelity with low error match.
DPO	OCEAN-STATIC	0.8676	0.0562	Highest DPO correctness.
DPO	OCEAN-DLS	0.8182	0.0507	Strong downstream fit but low same-error signal.
DPO	ATM27-STATIC	0.8654	0.0518	Strong acts/ROUGE with low same-error signal.
DPO	ATM27-DLS	0.7810	0.0606	Highest DPO same-error value but lower correctness.

Table 21: Mapped same-error examples. These are cases where the gold turn and prediction are both wrong and share a local math-error family.

Stage	Method	Dlg/turn	GT wrong turn	Pred wrong turn	Why it matters
Prompt	ATM27-STATIC	22/11	112?	100%	Both wrong with math tokens; useful same-error probe, not correctness.
Prompt	ATM27-STATIC	105/7	360?	180	Wrong numerical relation rather than arbitrary chat response.
Prompt	ATM27-STATIC	347/9	6.5	0.65	Decimal-scale misconception family.
SFT	NO-PERSONA	1/18	800	980	Compact wrong student-like answer after SFT.
SFT	ATM27-STATIC	24/9	15%	14.5%	Close percent misconception.
SFT	ATM27-STATIC	123/9	1/8	8/8	Fraction error family.
DPO	OCEAN-STATIC	1/18	800	980	Same example persists after DPO.
DPO	OCEAN-STATIC	11/12	0	0.2	Wrong numeric continuation.
DPO	ATM27-STATIC	53/5	10000	1000	Place-value/scale confusion.
DPO	ATM27-STATIC	84/5	3n?	2n	Algebraic coefficient error.
DPO	ATM27-STATIC	115/11	5?	12.5%	Percentage/answer-form mismatch.