

Simulating Multimodal Multiparty Student Interactions with Large Language Models for Training Adaptive Pedagogical Agents

Tristan Maidment, Timothy Nokes-Malach, Diane Litman,
Erin Walker, Adriana Kovashka

University of Pittsburgh

{tdm51, nokes, dlitman, eawalker}@pitt.edu, kovashka@cs.pitt.edu

Abstract

Multiparty tutoring policies must regulate turn-taking and social attention, yet collecting the interaction data needed for reinforcement learning is expensive in educational settings. We introduce an LLM-based user simulator for triadic robot–student dialogue that represents gaze and gesture as inline stage directions. Combined with phase-tagged prompting (which delineates conversational structure) and per-turn persona reinforcement, this design allows the LLMs to generate interactions that preserve key discourse structure and controllable participation dynamics.

We validate six simulator variants against a real student–robot corpus (12 student pairs) using metrics for semantic similarity, cue obedience, and persona consistency. While the simulator reaches strong semantic agreement with real transcripts, it also reveals a measurable sim-to-real style gap (in verbosity and grammatical complexity). The simulator enables the creation of an RL environment where the tutor selects factorized gaze and reference actions, learning to interact with simulated students. We compare online training (APPO) to offline learning (DQN). Off-policy evaluation on real logs shows that offline policies trained purely from synthetic data exceed online-trained policies under a doubly-robust estimator, and a verbal-only ablation yields unstable OPE due to heavy-tailed importance weights. These results support a scalable workflow for prototyping socially grounded multiparty tutoring policies while relaying where real-world coverage and OPE reliability remain bottlenecks.

1 Introduction

Developing dialogue policies to effectively manage multiparty interactions is a challenging task in computational linguistics (Sapkota et al., 2025; Ganesh et al., 2023). Unlike dyadic conversations, multiparty settings require agents to not

only process verbal content but also manage social dynamics, such as turn-taking regulation, gaze-based addressee selection, and non-verbal back-channeling (Castillo-López et al., 2025; Vertegaal and Ding, 2002; Ouchi and Tsuboi, 2016; Axelson et al., 2022). Reinforcement Learning (RL) offers a robust paradigm for optimizing such policies; yet its application is bottlenecked by the *data scarcity problem* (Kwan et al., 2023; Zhang et al., 2020). Training RL agents through direct interaction with humans—especially in sensitive domains like education—is prohibitively expensive, time-consuming, and fraught with ethical risks during the exploration phase (Memarian and Doleck, 2024; Akgun and Greenhow, 2022).

User simulation has long served as a surrogate to mitigate these costs (Young et al., 2013; Schatzmann et al., 2007). However, existing simulators often struggle to bridge the gap between linguistic fidelity and social realism. Traditional rule-based simulators lack linguistic diversity, while modern LLM-based approaches typically focus on semantic coherence at the expense of non-verbal dynamics (Luo et al., 2024). This omission should not be overlooked: in multiparty dialogue, *who* is looking at *whom* is often as important as what is being said (Vertegaal and Ding, 2002; Auer, 2018). A simulator that ignores these non-verbal cues may fail to effectively train an agent to manage participation or intervene at appropriate moments.

To address this gap, we present a novel framework for high-fidelity, multimodal user simulation using text-only Large Language Models. Our core insight is that complex social signals—gaze, gesture, and silence—can be effectively encoded as textual inline “stage directions” within the prompt context. This allows off-the-shelf LLMs to simulate rich, non-verbal multiparty dynamics without necessitating specialized multimodal architectures.

We ground this framework in the domain of adaptive pedagogical agents (Woolf, 2008), using

a dataset of student-robot interactions ($N = 12$ pairs) as our real-world benchmark. This specific context—collaborative math problem solving—provides a testbed requiring both task-oriented logic and subtle social coordination.

Our contributions are threefold:

1. **A Multiparty Simulation Framework:** We introduce a method for simulating multiparty interactions that integrates verbal and non-verbal channels via text, supported by our “Student Primitives” mechanism for controlling participation tendency. We validate this framework against real-world data, quantifying the “Sim-to-Real” gap, to estimate the social realism in generated dialogue (Da et al., 2025). We note that recent benchmarks in the math tutoring domain suggest that validating simulators against human judgment is critical for ensuring they serve as reliable proxies (Dou et al., 2025).
2. **The Necessity of Multimodal Grounding:** Through ablation studies, we provide empirical evidence that multimodal grounding is important for managing multiparty interactions using RL in our setting. We show that policies trained without gaze/gesture awareness suffer from distributional shift when evaluated against real-world human behavior, suggesting that non-verbal signals provide structural guidance for producing valid social interactions (Asis et al., 2023).
3. **Validating a Scalable Offline Workflow:** To circumvent the high latency and computational cost of live LLM interaction, we investigate offline reinforcement learning using synthetic logs. We demonstrate that policies trained offline achieve parity with those trained via expensive online simulation, establishing a high-throughput, low-resource framework for developing socially aware agents.

2 Simulation Design and Implementation

To enable the training of these RL policies, our simulation framework must accurately model not just the semantic content of student speech, but also the structural and non-verbal dynamics of multiparty interaction. We approach this as a *controlled generation* task, where an LLM is prompted to act as

a simulator for a structured educational discourse environment (Sun et al., 2023).

2.1 Target Discourse Dynamics

We model a triadic interaction involving a pedagogical robot (*Emma*) and two students. This setting is characterized by a shifting discourse structure: the dialogue oscillates between *Robot-Led* task steering and autonomous *Peer-to-Peer* discussion. Within these segments, key communicative acts—such as signaling confusion or claiming the floor—are mediated non-verbally through gaze and gesture. A successful simulator must reproduce this dynamic turn-governance to ensure the RL agent learns to navigate realistic social transitions.

2.2 Operationalizing Social Realism

We use the term *social realism* to refer to three measurable properties of multiparty interaction, rather than a holistic judgment of naturalness:

- *Gaze-Mediated Turn-Taking:* The student cued by the robot’s gaze or verbal reference should be the one who takes the next turn (Vertegaal and Ding, 2002; Auer, 2018).
- *Participation Balance:* Both students should contribute verbally. We treat balance as a proxy for the collaborative engagement described by the ICAP framework (Chi and Wylie, 2014), and operationalize it using word count entropy (Asano et al., 2024).
- *Initiative Asymmetry:* One student may consistently drive problem solving while the other follows. This pattern has been linked to knowledge co-construction in peer collaboration (Howard et al., 2017).

Section 3 evaluates the simulator on the first two properties directly (via the Phase Coherence Score and Word Count Entropy). The third motivates the Student Primitives mechanism (Section 2.6), which is designed to produce controllable asymmetry in initiative-taking.

2.3 Generative Simulation Framework

Our architecture (Figure 1) orchestrates the interaction between a standard LLM (functioning as the student simulator) and the RL agent. The robot’s verbal content is generated by a fixed rule-based manager using AIML (Artificial Intelligence Markup Language) (Wallace, 2009), while the RL

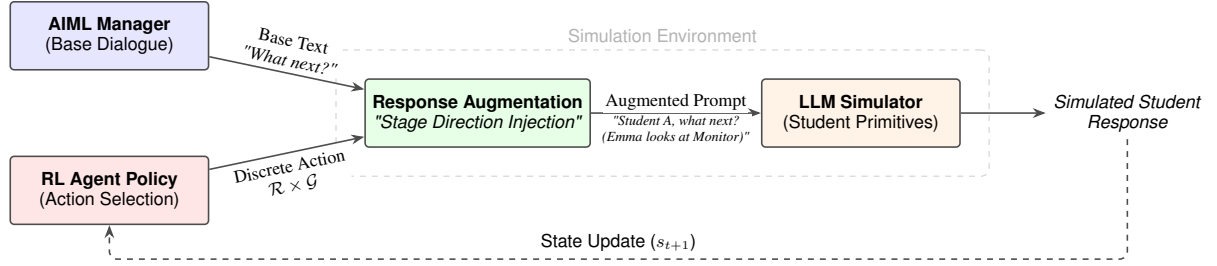


Figure 1: Core architecture of the LLM-based student simulation framework. Discrete RL actions (Gaze/Reference) are fused with rule-based dialogue (AIML) via the Response Augmentation module, creating a multimodal context for the LLM simulator.

agent modulates the social delivery (gaze and reference cues) of that content.

2.4 Textual Encoding of Multimodal Signals

To simulate non-verbal behaviors with a text-only LLM, we introduce a **Textual Encoding Mechanism** that translates multimodal signals into explicit inline “stage directions” embedded within the dialogue prompt.

This mechanism functions in two directions:

1. *Robot-to-Sim (Input)*: The RL agent’s actions (e.g., *Reference: Student A, Gaze: Monitor*) are converted into parenthetical descriptors appended to the robot’s utterance:

Emma: “What do you think is next?” (Emma looks at the monitor and shrugs)

This explicit encoding primes the LLM to condition its generation on the robot’s attention target, mimicking the effect of joint attention in human dialogue (Vertegaal and Ding, 2002; Tomasello, 2003).

2. *Sim-to-Robot (Output)*: We instruct the LLM to generate stage directions for the simulated students (e.g., (Student A looks at Student B)). These generated text tokens are parsed directly by the RL agent, allowing it to observe student gaze patterns as part of its state representation.

2.5 Structural Enforcement via Phases

Collaborative learning interactions in our domain follow a strict two-stage discourse structure: students first debate the answer amongst themselves (*Peer Discussion*) before presenting their consensus to the tutor (*Explanation*). This mirrors the deployed protocol, in which pairs solved each problem together before teaching the solution to Emma

(Appendix B.1). To enforce this sequence, we implement a **Phase Tagging** prompt strategy. We explicitly instruct the LLM to delineate its output using the following structural tags:

- [Phase 1]: Peer-to-peer discussion, where students debate the answer and are forbidden from addressing the robot.
- [Phase 2]: Explanation to the robot, where a single student must deliver the final answer to the tutor.

By forcing the model to generate these explicit tags, we impose a hard constraint on the dialogue flow, preventing the common failure mode where simulated students prematurely address the tutor or skip the collaborative step.

2.6 Persona Modeling and Consistency

A known limitation of LLMs in dialogue simulation is “persona drift” or “mode collapse,” where individual simulated characters gradually converge to a generic, average voice (Skantze, 2021). To counter this, we employ a **Student Primitives** framework. Each simulated student is assigned one of five discrete profiles, ranging from *Extremely Shy* to *Very Confident* (Appendix D). These tiers are a control knob for generating measurable variation in participation tendency, not a validated personality instrument. The construct we manipulate is *initiative-taking propensity* (how readily a student speaks first and holds the floor) rather than a psychological trait model. We make no claim that the tier labels correspond to psychological categories; our validation (Section 3) tests only whether the tiers produce distinct participation distributions. To maintain these traits over long interaction horizons, we propose **Dynamic Instruction Reinforcement (DIR)**; at every turn, we dynamically append a short, high-priority instruction *solely* to the prompt context,

reaffirming the specific traits of the current speakers (e.g., "*Student A is shy and rarely speaks first. Student B is dominant.*"). This additional prompt context is excluded from the dialogue history. This ephemeral reinforcement ensures that the generated participation patterns remain distinct and consistent throughout the episode, aiming to provide the RL agent with stable, varied social dynamics to learn from.

3 Simulation Validation

To validate our framework as a training ground for social policies, we assess its fidelity across three dimensions: Linguistic Realism (Do simulated students sound like real students?), Discourse Structure (Do they follow social norms?), and Personality Consistency (Do they maintain distinct personas?).

3.1 Reference Dataset

We ground our validation in an IRB-approved multimodal corpus of real-world student–robot interactions. The dataset consists of $N = 12$ undergraduate pairs (24 participants) engaging in collaborative math problem-solving with a robot tutor. The corpus comprises $N = 71$ conversations (one per math problem). Each conversation alternates between robot utterances and the student replies they elicit; we refer to one such exchange as a *turn*. The corpus contains 737 turns in total (10.4 ± 7.1 per conversation), and each turn corresponds to one state transition under the definition in Section 4.1. At the utterance level, the corpus contains 2,770 utterances (818 robot, 1,952 student; 39.0 ± 28.5 per conversation) averaging 16.0 ± 15.5 words (20.4 ± 9.7 for the robot, 14.1 ± 17.0 for students).

For comparison with the simulator, we aligned these streams into a linear dialogue history and converted the non-verbal signals into the inline stage-direction format of Section 2.4 (e.g., (*Student looks at Robot*)). The deployed protocol imposed the same solve-then-teach structure, so the phase tags in the real data reflect the collection protocol rather than post-hoc annotation. Details on the protocol, participants, and alignment pipeline are in Appendix B.

We compare six simulation configurations ($N = 600$ episodes) against our real-world corpus ($N = 71$ episodes), utilizing non-parametric statistical tests (Kolmogorov-Smirnov, Mann-Whitney U) to account for non-normal distributions in dialogue

data.

3.2 Experimental Configurations

We evaluate six simulation configurations to isolate the impact of our architectural components (Table 1). We test two LLM backbones: Qwen (Qwen3-14B-Chat) and Mistral (Mistral-Small-Instruct) (Yang et al., 2025; Mistral AI Team, 2025). For each, we perform ablation studies on our two mechanisms:

- *Dynamic Instruction Reinforcement (DIR)*: The per-turn injection of personality traits.
- *Dynamic Temperature Management (DTM)*: A simple adaptive sampling strategy that adjusts generation temperature based on validation success rates. (Appendix E.2)

Table 1: Simulation Configurations. DIR: Dynamic Instruction Reinforcement; DTM: Dynamic Temperature Management.

Configuration	LLM Backbone	DIR	DTM
Qwen-Full	Qwen3-14B	✓	✓
Qwen-NoDIR	Qwen3-14B		✓
Mistral-Full-DIR	Mistral-24B	✓	✓
Mistral-NoDIR	Mistral-24B		✓
Mistral-Full-NoTemp	Mistral-24B	✓	
Mistral-BaseLLM	Mistral-24B		

3.3 Dimension 1: Linguistic Realism

We evaluate the semantic and stylistic quality of generated utterances using existing NLP metrics.

Semantic Coherence: We verify that simulated responses are contextually relevant using BERTScore, measuring the similarity between simulated utterances and the distribution of real-world responses in similar contexts (Zhang et al., 2019). As shown in Table 2, all primary configurations achieve high F1 scores (> 0.86), outperforming a randomized text baseline (Lorem Ipsum, 0.62) which establishes the lower bound for non-semantic overlap. More importantly, the model approaches the *Inter-Speaker Consistency* score (0.901), which quantifies the semantic agreement between actual human pairs and serves as a ceiling for naturalistic variation.

Table 5 provides an excerpt from a simulated interaction, illustrating the interplay between phase tags, stage directions, and personality expression.

Stylistic "Sim-to-Real" Gap: While semantically coherent, we observe a stylistic gap. Analysis of

Table 2: BERTScore F1 Results Comparing Simulated Dialogues to Real-World Corpus. Higher F1 scores indicate greater semantic similarity. “Overall” refers to the complete dialogue; “Phase 1” to Peer Discussion; and “Phase 2” to Explanation to Robot.

Simulation Configuration	Overall F1	Phase 1 F1	Phase 2 F1
Qwen-Full	0.8741	0.8943	0.8728
Qwen-NoDIR	0.8519	0.8415	0.8683
Mistral-Full-DIR	0.8626	0.8600	0.8747
Mistral-NoDIR	0.8594	0.8415	0.8787
Mistral-Full-NoTemp	0.8631	0.8510	0.8812
Mistral-BaseLLM	0.8591	0.8451	0.8798
Lorem Ipsum Baseline	0.6229	0.6175	0.6282
Real World Inter-Speaker Consistency	0.9019	0.9223	0.9093

linguistic complexity metrics reveals that simulated students are consistently more verbose and grammatically complex than their real-world counterparts.

- *Verbosity*: Simulated utterances have a higher Average Sentence Length (ASL) ($p < 0.001$, Cohen’s $d \approx 2.4$).
- *Complexity*: Flesch-Kincaid Grade Level (FKGL) is higher in simulation ($M = 6.2$) than in real transcripts ($M = 4.1$) (Kincaid et al., 1975).

3.4 Dimension 2: Discourse Structure & Responsiveness

One requirement for our simulator is that agents respect the specific turn-taking norms of the peer-to-peer (Peer) and human-robot (Robot) phases. We define a metric, the **Phase Coherence Score (PCS)**, to quantify this:

$$PCS = \frac{1}{N} \sum_{t \in T} \mathbb{I}(\text{speaker}_t = \text{target}_t) \quad (1)$$

where target_t is the student explicitly cued by Emma’s gaze/reference. A score of 1.0 implies perfect obedience to social cues.

Result: Variable Social Compliance. PCS is defined only for episodes containing at least one student-to-robot initiation, so we report it alongside the fraction of episodes in which such initiations occur (all 71 real conversations). We compare each configuration to the real corpus with two-sided Mann-Whitney U tests (rank-biserial r as effect size) and report bootstrapped 95% confidence intervals (10,000 resamples) on the mean. Full statistics for all configurations are in Appendix C. Compliance was highly model-dependent:

Qwen-Full students addressed the robot in 84% of episodes, but with low coherence (PCS $M = 0.24$, 95% CI $[0.21, 0.27]$), significantly below real students ($M = 0.37$, $[0.30, 0.43]$; $U = 6433.5$, $n_1 = 253$, $n_2 = 71$, $p < 0.001$, $r = -0.28$). This creates a "chaotic" training environment where students frequently ignore the robot’s cues.

Mistral-Full-DIR students addressed the robot in fewer episodes (49%), but when they did, no significant difference from real students was detected (PCS $M = 0.44$, $[0.38, 0.50]$; $U = 5696.5$, $n_1 = 146$, $n_2 = 71$, $p = 0.23$, $r = 0.10$). This null result does not establish equivalence; the intervals are consistent with small differences in either direction.

The gap between the two backbones is itself useful: it allows us to "tune" the difficulty of the social environment. We can train agents in the more compliant Mistral environment for predictability, or in the "chaotic" Qwen environment for robustness.

3.5 Dimension 3: Personality Consistency

Finally, we validate that our "Student Primitives" effectively modulate participation. We define **Word Count Entropy (WCE)** as a measure of participation balance across students (1.0 = perfect equity) (Shannon, 1948; Asano et al., 2024).

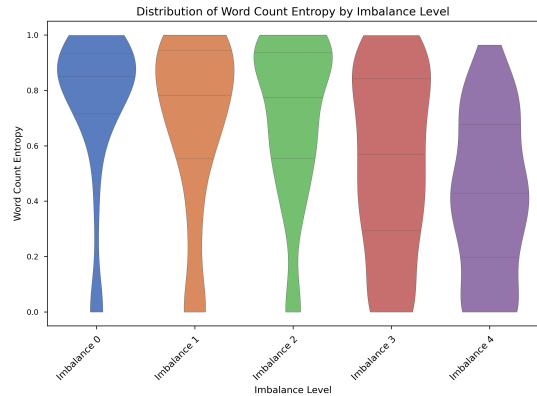


Figure 2: Distribution of Word Count Entropy (WCE) across simulated personality imbalance levels (0 to 4). WCE decreases with the assigned imbalance, the signature expected if the “Student Primitives” control participation as intended.

Because the Imbalance Level is the absolute difference between the two students’ assigned participation-tendency scores (Appendix D), a working control mechanism should produce a specific signature: as the assigned gap grows, the more confident student should take a larger share of the

words, and WCE should fall. This is what we observe (Figure 2): mean WCE decreases from 0.94 for balanced pairs (Level 0) to 0.58 for extreme imbalance (Level 4), with intermediate levels between these extremes. The negative relationship is the expected outcome of the prompts having their intended effect; a flat trend would indicate that the personality descriptions were being ignored.

Result: Dynamic Instruction Reinforcement (DIR) is required to sustain this effect. A Kruskal-Wallis test confirms that the five imbalance levels produce distinct WCE distributions ($H(4) = 77.8$, $p < 0.001$). Our ablation study reveals that without per-turn reinforcement, the influence of the initial system prompt decays as the context window grows (context dilution). This leads to distinct personalities (e.g., Shy vs. Confident) gradually converging to a generic average behavior, such that no significant differences between imbalance levels remain detectable over long episodes.

4 Reinforcement Learning Environment Design

With the simulation validation complete, we formulated the RL problem to train policies for managing student participation. We treat the LLM as a generator of environment dynamics rather than the policy itself, as standard language models cannot directly optimize for non-differentiable pedagogical rewards (such as participation entropy or learning gains). RL enables us to explicitly maximize these specific educational outcomes. We model the task as a Partially Observable Markov Decision Process (POMDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$ (Kaelbling et al., 1998).

4.1 State and Action Space

State Space ($\mathcal{S}$): Unlike traditional approaches that rely on hand-crafted features (D’Mello et al., 2007; Forbes-Riley and Litman, 2011; Chi et al., 2011; Pietquin et al., 2011), we utilize a raw textual representation (Jaques et al., 2019). The state s_t at turn t (one robot utterance and the student replies it elicits, as in Section 3.1) consists of the tokenized dialogue history $H_t = [u_0, u_1, \dots, u_t]$, where each utterance u_i includes both verbal content and the inline stage directions defined in Section 2.4. To maintain computational tractability, we employ a sliding context window of the most recent K turns.

Action Space ($\mathcal{A}$): The agent controls Emma through a factorized discrete action space

$\mathcal{A} = \mathcal{R} \times \mathcal{G}$ ($|\mathcal{A}| = 12$). This consists of Verbal Reference $\mathcal{R} \in \{\text{Student A, Student B, None}\}$ (determining if the robot explicitly names a student) and Gaze Direction $\mathcal{G} \in \{\text{Student A, Student B, Monitor, Thinking}\}$ (setting the non-verbal attention target). To ensure naturalism, we apply action masking to disable the verbal reference component $\mathcal{R}$ when the base dialogue system is not generating a question.

4.2 Reward Function

The reward function $R_{\text{total}}(t) = \alpha_E R_E(t) + \alpha_D R_D(t)$ (with $\alpha_E = \alpha_D = 0.5$) combines two components grounded in the ICAP framework (Chi and Wylie, 2014): Entropy-based Balance (R_E), which measures the evenness of verbal participation using the Shannon entropy of word counts; and Dialogue Expressivity Ratio (R_D), which measures lexical alignment between students via shared unique expressions (Dubuisson Duplessis et al., 2021; Friedberg et al., 2012).

4.3 Optimization Algorithm

Due to the computationally intensive nature of LLM-based simulation (step times often exceeding 30s), traditional on-policy algorithms like PPO are prohibitively slow. We instead employ **Asynchronous PPO (APPO)**, an implementation of the IMPALA architecture (Espeholt et al., 2018). APPO decouples data collection (Actors) from optimization (Learner), allowing us to parallelize simulation across multiple vLLM (Kwon et al., 2023) workers while using IMPALA’s V-trace off-policy corrections to handle the lag between behavior and target policies.

5 RL Findings and Analysis

We trained several RL policies varying the algorithm (APPO vs PPO) and the training data distribution. To assess robustness across different contexts, we employed two personality sampling strategies:

Uniform: Individual student traits are sampled uniformly at random, resulting in a normal distribution of imbalance centered on moderate pairs.

Imbalance-Focused: The sampling probability is heavily skewed towards high-imbalance dyads (e.g., Level 3 and 4), deliberately exposing the agent to challenging scenarios where participation is naturally skewed.

5.1 Effectiveness of RL Training

We analyze the performance of seven policy configurations to isolate the effects of algorithm choice, training distribution, and data source. These include online agents varying by algorithm (APPO vs. PPO) and sampling curriculum (Uniform vs. ImbalanceFocused), as well as a high-entropy variant (0.1 Ent) designed to be similar to the behavior of the Random baseline. We also introduce a multimodal ablation (No Gaze/Gesture), trained in an environment where the LLM is neither prompted with nor generates non-verbal “stage directions,” strictly relying on verbal tokens. Additionally, we evaluate offline agents trained via DQN on datasets of varying composition: purely synthetic (Offline-Simulated), purely real-world (Offline-Real), and a 50/50 mix (Offline-Mixed). Table 3 summarizes their performance.

Table 3: Comparative Performance of Policy Ablations in Simulation. We evaluate policies based on In-Simulation Reward (Mean Episode Reward) using Uniform sampling. Policies training used a default entropy of 0.05 unless specified. Italicized rows denote reference policies that serve as theoretical upper bounds.

Policy	Sampling	Algorithm	Sim Reward
Random (Baseline)	N/A	N/A	4.8713 $\pm$ 2.6707
PPO	Uniform	PPO	4.3401 $\pm$ 2.8202
Uniform	Uniform	APPO	4.7386 $\pm$ 3.1636
Offline-Simulated	Uniform	DQN	4.7927 $\pm$ 3.1943
ImbalanceFocused	Imbalanced	APPO	5.1612 $\pm$ 3.1446
ImbalanceFocused (0.1 Ent)	Imbalanced	APPO	5.2605 $\pm$ 3.5395
ImbalanceFocused (No Gaze/Gesture)	Imbalanced	APPO	5.1299 $\pm$ 2.7334
<i>Offline-Mixed (Reference)</i>	<i>Uniform</i>	<i>DQN</i>	5.5675 $\pm$ 2.976
<i>Offline-Real (Reference)</i>	<i>Uniform</i>	<i>DQN</i>	4.8101 $\pm$ 3.3901

Algorithmic Efficiency (APPO vs. PPO): We observe a performance gap between asynchronous and synchronous methods. The ImbalanceFocused (APPO) agent achieved a mean reward of 5.16, outperforming the PPO agent (4.34). Both were trained for the same *wall-clock duration*, so the gap reflects APPO’s throughput advantage in high-latency LLM environments (9.97 vs. 1.46 aggregate steps/sec), where decoupling simulation from learning improves sample efficiency

Curriculum and Generalization: The sampling strategy used during training had a marked effect on final performance. Agents trained on ImbalanceFocused distributions (5.16) outperformed those trained on Uniform distributions (4.74) and the PPO baseline (4.34), *even when evaluated on Uniform scenarios*. This suggests that exposing agents to “hard” cases (high participation imbalance) forces the policy to learn more robust inter-

vention strategies than training on the target distribution itself. While the Random baseline (4.87) remains competitive by artificially inflating the entropy reward (R_E), the RL agent achieves its higher score by balancing R_E with Dialogue Expressivity (R_D), ensuring that the interaction is coherent rather than merely random.

Emergent Social Strategies: The agent moved beyond stochastic behavior to adopt distinct social strategies. Qualitative analysis of high-advantage states reveals observable learned behaviors:

1. **Congruent Cueing for Engagement:** When introducing new problems, the policy learned to couple verbal references with direct gaze (e.g., *Reference: Student A, Gaze: Student A*), even though the action space permits incongruent combinations to support deictic gesturing (Appendix E.1).
2. **Identifying Unproductive States:** The agent assigned lower advantage to interventions during system artifacts (e.g., repetitive AIML glitches such as “My body looks like unknown”), recognizing when standard strategies were failing to advance the state.

5.2 The Utility of Synthetic Data

Evaluation Methodology. Deploying every candidate policy with real students is infeasible, so we estimate the value of a target policy π from trajectories logged under a different behavior policy π_b (here, the randomized gaze and reference policy used during the original deployment; Appendix B.1). This is the Off-Policy Evaluation (OPE) problem. We report four standard estimators that occupy different points on the bias–variance spectrum (Voloshin et al., 2021):

- **Importance Sampling (V_{IS}):** Reweights each logged return by the likelihood ratio $\prod_t \pi(a_t | s_t) / \pi_b(a_t | s_t)$ (Precup et al., 2000). Unbiased, but its variance grows exponentially with the episode horizon.
- **Weighted IS (V_{WIS}):** Normalizes the importance weights across trajectories (Precup et al., 2000). This introduces a small bias in exchange for substantially lower variance.
- **Direct Method (V_{DM}):** Fits a regression model of the value function from the logged data and evaluates π under that model (Le et al., 2019). Low variance, but biased whenever the model is misspecified.

- *Doubly Robust (V_{DR})*: Combines V_{DM} and V_{IS} , using the model as a control variate for the importance-weighted returns (Jiang and Li, 2016; Thomas and Brunskill, 2016). It is unbiased if *either* the value model or the importance weights are accurate, and typically has lower variance than V_{IS} .

We designate V_{DR} as our primary estimator. Our reference corpus is small ($N = 71$ episodes) and the horizon is long ($H \approx 16$), conditions under which pure importance weights become heavy-tailed while V_{DM} alone is vulnerable to model bias. We report all four estimators for transparency and treat disagreement among them as a diagnostic of weak overlap between π and π_b (Section 5.3).

A central question is whether we can leverage synthetic data for efficient policy learning. We compared an Online agent (trained via live APPO interaction with the simulator) against an Offline-Simulated agent (trained via DQN on a dataset of random simulator interactions, $N = 6,038,245$).

To contextualize these results, we also trained two reference policies using our limited real-world corpus ($N = 71$ episodes):

- *Offline-Real (Upper Bound)*: Trained solely on the real-world corpus. We acknowledge that evaluating on the training set is non-standard; however, this is intended to serve as a *generous theoretical ceiling*, representing an "oracle" policy that has effectively memorized the optimal actions for this specific target distribution.
- *Offline-Mixed (Augmented)*: Trained on a 50/50 mix of real and synthetic data.

Table 4 presents OPE estimates for each policy on the real-world corpus (Asis et al., 2023).

Synthetic Data Viability and Data Limitations: The Offline-Simulated agent ($V_{DR} = 1.009$) and Online agent (1.012) perform comparably to the Random baseline (0.990), with overlapping confidence intervals. While this parity suggests the policies are "safe" (not degrading performance), the lack of statistical separation highlights a critical limitation: our small real-world dataset ($N = 71$) provides insufficient statistical power to reliably distinguish between competent policies using OPE. However, the trends are directionally consistent—simulation-trained agents score nominally higher than random, and the Offline-Real agent scores

highest (1.126). This score establishes a theoretical performance ceiling and confirms the sensitivity of our metric. Notably, the Offline-Mixed result (1.056) falls between the synthetic and real-only policies, suggesting that while synthetic data is useful for initialization, naive mixing can dilute the high-quality signal of real experiences.

5.3 The Value of Multimodal Grounding

To assess the value of "stage direction" encoding, we evaluated an ablated policy (No Gaze/Gesture) that observed only verbal tokens. While this agent performed well within the simulation (Reward: 5.13), its OPE evaluation on real-world data revealed an interesting limitation.

Specifically, the OPE process entered a weak-overlap (limited effective support) regime. Although the data-collection policy was uniform random over actions (so support holds in principle), the No Gaze/Gesture policy converged to a substantially more peaked action distribution. Under trajectory importance sampling, the probability that a random behavior trajectory matches such a specific policy decays exponentially with episode length ($H \approx 16$). This yields a heavy-tailed distribution of importance weights (Thomas and Brunskill, 2016) where the Effective Sample Size (ESS) collapses to near zero. Consequently, the OPE estimate becomes dominated by a single "lucky" trajectory that happened to match the policy, rendering the resulting value estimate statistically meaningless and highly unstable.

This mismatch manifested numerically as exploded trajectory weights: standard V_{IS} estimates exceeded 1.0×10^6 , and the effective sample size (ESS) collapsed to < 2.0 , indicating that the estimate was dominated by a small number of trajectories (Metelli et al., 2020).

6 Related Work

Multiparty Dialogue and Turn-Taking. Most dialogue systems research targets dyadic settings; multiparty interaction adds addressee selection, floor management, and speaker identification (Ganesh et al., 2023; Sapkota et al., 2025). Gaze is a primary cue for these decisions in three-party conversation (Vertegaal and Ding, 2002; Auer, 2018), but computational models of addressee selection and turn-taking (Ouchi and Tsuboi, 2016; Skantze, 2021; Castillo-López et al., 2025) have largely treated it as a prediction target. We instead treat gaze and ver-

Table 4: Off-Policy Evaluation (OPE) Results on Real-World Data. We report estimates from four estimators: Direct Method (V_{DM}), Doubly Robust (V_{DR}), Weighted Importance Sampling (V_{WIS}), and Importance Sampling (V_{IS}). Values represent the mean estimate with 95% confidence intervals in parentheses. V_{DR} is bolded as the primary metric due to its balance of bias and variance.

Policy	V_{DM} (Model-Based)	V_{DR} (Doubly Robust)	V_{WIS} (Weighted IS)	V_{IS} (Std. IS)
Random (Baseline)	0.972 (0.724–1.197)	0.990 (0.945–1.036)	1.001 (0.973–1.036)	0.999 (0.967–1.026)
PPO (Online)	0.896 (0.666–1.208)	0.921 (0.673–1.223)	0.929 (0.675–1.129)	0.914 (0.674–1.172)
Uniform (Online)	1.070 (0.798–1.313)	0.933 (0.583–1.214)	0.966 (0.832–1.158)	0.966 (0.823–1.147)
Offline-Simulated	1.334 (0.993–1.646)	1.009 (0.936–1.117)	1.002 (0.986–1.050)	1.002 (0.979–1.052)
ImbalanceFocused (Online)	1.070 (0.795–1.304)	1.012 (0.489–1.456)	0.952 (0.717–1.166)	0.953 (0.709–1.166)
Imb. Focused (0.1 Ent)	1.197 (0.903–1.484)	1.027 (0.550–1.395)	0.960 (0.768–1.097)	0.952 (0.762–1.108)
No Gaze/Gesture	<i>Unreliable**</i>	<i>Unreliable**</i>	<i>Unreliable**</i>	<i>Unreliable**</i>
<i>Offline-Mixed (Reference)</i>	<i>1.051 (0.782–1.294)</i>	<i>1.056 (0.931–1.162)</i>	<i>0.986 (0.902–1.060)</i>	<i>0.988 (0.906–1.075)</i>
<i>Offline-Real (Reference)</i>	<i>0.943 (0.708–1.167)</i>	<i>1.126 (0.899–1.414)</i>	<i>0.943 (0.802–1.092)</i>	<i>0.914 (0.707–1.103)</i>

** Note: OPE estimates for the ‘No Gaze/Gesture’ policy were consistently extremely large ($> 10^6$), likely due to severe off-policy distribution shift, rendering them unreliable for interpretation.

bal reference as controllable actions whose effects on student behavior the RL agent must learn.

User Simulation. Agenda-based and statistical simulators enabled early POMDP dialogue policy learning (Schatzmann et al., 2007; Young et al., 2013), but lack linguistic diversity. LLM-based simulators restore that diversity (Luo et al., 2024; Sun et al., 2023), although their reliability as proxies for human users remains an open concern (Dou et al., 2025), and they are almost exclusively dyadic and text-only. We extend them to triadic interaction with textually encoded non-verbal channels, validated against a real corpus.

Reinforcement Learning for Tutoring. RL has been used to induce pedagogical policies from tutoring corpora (Chi et al., 2011; D’Mello et al., 2007; Forbes-Riley and Litman, 2011; Memarian and Doleck, 2024), typically with hand-crafted state features and data from a single learner. Closest to our setting, Maidment et al. (2022) built an RL environment for a teachable robot from a limited multiparty corpus; we differ in using an LLM simulator to generate unbounded training interactions and in learning from raw text states (Jaques et al., 2019). Our reward draws on the ICAP framework (Chi and Wylie, 2014), on participation-balance metrics that predict learning with a robot (Asano et al., 2024), on initiative shifts as a signal of knowledge co-construction (Howard et al., 2017), and on lexical entrainment as a predictor of group success (Friedberg et al., 2012; Dubuisson Duplessis et al., 2021).

Sim-to-Real Transfer and Off-Policy Evaluation. Sim-to-real transfer is well studied in robotics

(Zhao et al., 2020; Da et al., 2025), where the gap is physical; in dialogue the gap is linguistic and social. Off-policy evaluation makes it possible to assess simulator-trained policies on real logs without deployment (Precup et al., 2000; Jiang and Li, 2016; Thomas and Brunskill, 2016; Voloshin et al., 2021), and has been applied to dialogue policies (Jaques et al., 2019; Gao et al., 2023). We use OPE both to compare online and offline training and, through its failure on the verbal-only ablation, as a diagnostic of distributional shift.

7 Conclusion

This work introduces a framework for training multiparty pedagogical policies using LLM-based multimodal user simulation. By encoding gaze, gesture, and social signals as textual “stage directions,” we enable text-only models to function as surrogates for these social interactions. Our results support three conclusions. *Multimodal grounding matters:* the No Gaze/Gesture ablation could not be evaluated reliably with OPE, indicating that the textual encoding aligns the simulator’s state distribution with real-world dynamics. *Personality-driven curricula help:* agents trained on ImbalanceFocused distributions ($R = 5.16$) outperformed those trained on uniform distributions ($R = 4.74$), suggesting that exposure to distinct participation profiles yields more robust strategies. *Offline learning scales:* policies trained purely on synthetic logs reached parity ($V_{DR} \approx 1.01$) with computationally expensive online agents, allowing large synthetic datasets to replace live simulation as the bottleneck.

8 Limitations

While our framework demonstrates the efficacy of LLM-based simulation for training pedagogical policies, we acknowledge several limitations that constrain the generalizability and interpretation of our results.

Statistical Power of Real-World Benchmarks

A primary bottleneck in our validation pipeline is the size of the real-world reference corpus ($N = 71$ conversations). While sufficient for establishing qualitative baselines, this small sample size limits the statistical resolution of our Off-Policy Evaluation (OPE). As seen in Table 4, the confidence intervals for competent policies (e.g., *Offline-Simulated* vs. *Online*) often overlap. Consequently, we interpret our OPE results as providing *directional evidence* of policy viability rather than a definitive ordinal ranking. Future work must prioritize larger-scale data collection to enable high-confidence policy separation.

Sim-to-Real Stylistic Gap Although our simulator achieves high semantic coherence (BERTScore > 0.86), we observe a measurable stylistic gap between simulated and real students. As detailed in Section 3, simulated students exhibit higher verbosity (Average Sentence Length) and grammatical complexity (Flesch-Kincaid Grade Level) than their real-world counterparts. This "verbosity bias" is a known artifact of instruction-tuned LLMs (Saito et al., 2023). There is a risk that RL agents trained in this environment may be operating out-of-distribution, as they were optimized for verbose interlocutors and may require calibration when deployed with more concise real-world students.

OPE Stability and Distributional Shift Our experiments highlight the fragility of Importance Sampling-based OPE estimators when the target policy deviates significantly from the behavior policy. The failure of the *No Gaze/Gesture* ablation (which generated importance weights $> 10^6$) underscores a known challenge in offline RL: when a policy learns a highly peaked or deterministic strategy that differs from the stochastic data-logging policy, the Effective Sample Size (ESS) collapses, rendering evaluation impossible. This limits our ability to evaluate these policy shifts without collecting new on-policy real-world data.

Domain and Embodiment Specificity Our simulator is currently grounded in a specific triadic

math tutoring scenario with a specific robotic embodiment (*Emma*). The "Student Primitives" and prompt structures are designed for collaborative problem-solving. It remains an open question how well this *textual encoding* approach generalizes to highly physical tasks (e.g., collaborative assembly) where non-verbal signals are continuous and spatial rather than discrete communicative acts.

Ethical Considerations and Deployment Risks

Our framework optimizes for participation balance and lexical alignment, which are proxy metrics that do not capture the full complexity of effective pedagogy. An agent trained on these objectives could, for example, learn to over-prompt reticent students in ways that increase verbal participation but cause discomfort or disengagement. As discussed in Section 2.6, the "Student Primitives" labels are a simulation control and a simplification of real student behavior; they should not be applied to real learners. More broadly, deploying RL-trained policies with real students without first closing the sim-to-real gap (Section 3) carries inherent risk, as unexpected agent behaviors in sensitive educational contexts could negatively affect the learning experience. Any future classroom deployment should incorporate human oversight and safety evaluation.

Acknowledgments

This work was supported by Grant No. 2024645 from the National Science Foundation, Grant No. 220020483 from the James S. McDonnell Foundation, and a University of Pittsburgh Learning Research and Development Center internal award. We thank the members of the PETAL lab and FacetLab at the University of Pittsburgh, and the anonymous reviewers for their thoughtful comments. Generative AI tools were used as writing and coding assistants throughout the preparation of this paper and its accompanying code; the research ideas, experimental design, and analyses are the authors' own, and all AI-assisted text and code were reviewed and verified by the authors.

References

- Selin Akgun and Christine Greenhow. 2022. [Artificial intelligence in education: Addressing ethical challenges in K-12 settings](#). *Ai and Ethics*, 2(3):431–440.
- Yuya Asano, Diane Litman, Quentin King-Shepard, Tristan Maidment, Tyree Langley, Teresa Davison, Timothy Nokes-Malach, Adriana Kovashka, and Erin

- Walker. 2024. What metrics of participation balance predict outcomes of collaborative learning with a robot? In *Educational Data Mining (EDM)*.
- Kristopher De Asis, Eric Graves, and Richard S. Sutton. 2023. [Value-aware Importance Weighting for Off-policy Reinforcement Learning](#). In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, pages 745–763. PMLR.
- Peter Auer. 2018. [Chapter 9. Gaze, addressee selection and turn-taking in three-party interaction](#). In Geert Br ne and Bert Oben, editors, *Eye-tracking in Interaction: Studies on the role of eye gaze in dialogue*, Advances in Interaction Studies, pages 197–232. John Benjamins Publishing Company.
- Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. [Modeling Feedback in Interaction With Conversational Agents—A Review](#). *Frontiers in Computer Science*, 4.
- Galo Castillo-L pez, Gael de Chalendar, and Nasredine Semmar. 2025. [A Survey of Recent Advances on Turn-taking Modeling in Spoken Dialogue Systems](#). In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 254–271, Bilbao, Spain. Association for Computational Linguistics.
- Mich lene T. H. Chi and Ruth Wylie. 2014. [The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes](#). *Educational Psychologist*, 49(4):219–243.
- Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. 2011. [Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies](#). *User Modeling and User-Adapted Interaction*, 21(1-2):137–180.
- Longchao Da, Justin Turnau, Thirulogasankar Pranav Kutralingam, Alvaro Velasquez, Paulo Shakarian, and Hua Wei. 2025. [A Survey of Sim-to-Real Methods in RL: Progress, Prospects and Challenges with Foundation Models](#). *arXiv preprint*. ArXiv:2502.13187 [cs].
- Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. 2025. [SimulatorArena: Are User Simulators Reliable Proxies for Multi-Turn Evaluation of AI Assistants?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35212–35290, Suzhou, China. Association for Computational Linguistics.
- Guillaume Dubuisson Duplessis, Caroline Langlet, Chlo  Clavel, and Fr d ric Landragin. 2021. [Towards alignment strategies in human-agent interactions based on measures of lexical repetitions](#). *Language Resources and Evaluation*, 55(2):353–388.
- Sidney D’Mello, Rosalind W. Picard, and Arthur C. Graesser. 2007. [Toward an affect-sensitive autotutor](#). *IEEE Intelligent Systems*, 22(4):53–61.
- Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. 2018. [IM-PALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures](#). *arXiv preprint*. ArXiv:1802.01561 [cs].
- Kate Forbes-Riley and Diane Litman. 2011. [Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system](#). *Computer Speech & Language*, 25(1):105–126.
- Heather Friedberg, Diane Litman, and Susannah B. F. Paletz. 2012. [Lexical entrainment and success in student engineering groups](#). In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 404–409.
- Ananya Ganesh, Martha Palmer, and Katharina Kann. 2023. [A Survey of Challenges and Methods in the Computational Modeling of Multi-Party Dialog](#). In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154, Toronto, Canada. Association for Computational Linguistics.
- Qitong Gao, Ge Gao, Juncheng Dong, Vahid Tarokh, Min Chi, and Miroslav Pajic. 2023. [Off-policy evaluation for human feedback](#). *arXiv preprint* arXiv:2310.07123.
- Cynthia Howard, Barbara Di Eugenio, Pamela Jordan, and Sandra Katz. 2017. [Exploring Initiative as a Signal of Knowledge Co-Construction During Collaborative Problem Solving](#). *Cognitive Science*, 41(6):1422–1449.
- Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson,  gata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. [Way off-policy batch deep reinforcement learning of implicit human preferences in dialog](#). In *Proceedings of the 36th International Conference on Machine Learning*.
- Nan Jiang and Lihong Li. 2016. [Doubly Robust Off-policy Value Evaluation for Reinforcement Learning](#). In *Proceedings of The 33rd International Conference on Machine Learning*, pages 652–661. PMLR.
- Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. 1998. [Planning and acting in partially observable stochastic domains](#). *Artificial Intelligence*, 101(1):99–134.
- J. P. Kincaid, Jr. Fishburne, Rogers Robert P., Chissom Richard L., and Brad S. 1975. [Derivation of New Readability Formulas \(Automated Readability Index, Fog Count and Flesch Reading Ease Formula\) for Navy Enlisted Personnel](#). Technical report, Defense Technical Information Center, Fort Belvoir, VA.
- Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. [A Survey on Recent Advances and Challenges in Reinforcement Learning Methods for Task-oriented Dialogue Policy Learning](#). *Machine Intelligence Research*, 20(3):318–334.

- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient Memory Management for Large Language Model Serving with PagedAttention](#). *arXiv preprint*. ArXiv:2309.06180 [cs].
- Hoang M. Le, Cameron Voloshin, and Yisong Yue. 2019. [Batch Policy Learning under Constraints](#). *arXiv preprint*. ArXiv:1903.08738 [cs, math, stat].
- Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. 2024. [DuetSim: Building User Simulator with Dual Large Language Models for Task-Oriented Dialogues](#). *arXiv preprint*. ArXiv:2405.13028 [cs].
- Tristan Maidment, Mingzhi Yu, Nikki Lobczowski, Adriana Kovashka, Erin Walker, Diane Litman, and Timothy Nokes-Malach. 2022. Building a reinforcement learning environment from limited data to optimize teachable robot interventions. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 62–73.
- Bahar Memarian and Tenzin Doleck. 2024. [A scoping review of reinforcement learning in education](#). *Computers and Education Open*, 6:100175.
- Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. 2020. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21(1):141:5552–141:5626.
- Mistral AI Team. 2025. [Mistral-small-24b-instruct-2501](#). Hugging Face model card. Accessed 2025-12-14.
- Hiroki Ouchi and Yuta Tsuboi. 2016. [Addressee and Response Selection for Multi-Party Conversation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143, Austin, Texas. Association for Computational Linguistics.
- Olivier Pietquin, Matthieu Geist, Senthilkumar Chandramohan, and Hervé Frezza-Buet. 2011. [Sample-efficient batch reinforcement learning for dialogue management optimization](#). *ACM Transactions on Speech and Language Processing*, 7(3):7:1–7:21.
- Doina Precup, Richard S. Sutton, and Satinder Singh. 2000. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 759–766, Stanford, CA.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust Speech Recognition via Large-Scale Weak Supervision](#). *arXiv preprint*. ArXiv:2212.04356 [eess].
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. [Verbosity Bias in Preference Labeling by Large Language Models](#). *arXiv preprint*. ArXiv:2310.10076 [cs].
- Sagar Sapkota, Mohammad Saqib Hasan, Mubarak Shah, and Santu Karmaker. 2025. [Multi-Party Conversational Agents: A Survey](#). *arXiv preprint*. ArXiv:2505.18845 [cs].
- Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, NAACL-Short '07, pages 149–152, USA. Association for Computational Linguistics.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Gabriel Skantze. 2021. [Turn-taking in Conversational Systems and Human-Robot Interaction: A Review](#). *Computer Speech & Language*, 67:101178.
- Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. [Evaluating Large Language Models on Controlled Generation Tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.
- Philip S. Thomas and Emma Brunskill. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, pages 2139–2148, New York, NY, USA. JMLR.org.
- Michael Tomasello. 2003. [Constructing a Language: A Usage-Based Theory of Language Acquisition](#). Harvard University Press.
- Roel Vertegaal and Yaping Ding. 2002. [Explaining effects of eye gaze on mediated group conversations: amount or synchronization?](#) In *Proceedings of the 2002 ACM conference on Computer supported cooperative work, CSCW '02*, pages 41–48, New York, NY, USA. Association for Computing Machinery.
- Cameron Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. 2021. [Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning](#). *arXiv preprint*. ArXiv:1911.06854 [cs, stat].
- Richard S. Wallace. 2009. [The Anatomy of A.L.I.C.E.](#) In Robert Epstein, Gary Roberts, and Grace Beber, editors, *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, pages 181–210. Springer Netherlands, Dordrecht.
- Beverly Woolf. 2008. *Building Intelligent Interactive Tutors, Student-Centered Strategies for Revolutionizing E-Learning*.

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](#). *arXiv preprint*. ArXiv:2505.09388 [cs].

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. [POMDP-Based Statistical Spoken Dialog Systems: A Review](#). *Proceedings of the IEEE*, 101(5):1160–1179.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [BERTScore: Evaluating Text Generation with BERT](#). *ArXiv*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. [Recent advances and challenges in task-oriented dialog systems](#). *Science China Technological Sciences*, 63(10):2011–2027.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. 2020. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. *arXiv preprint arXiv:2009.13303*. Published in IEEE SSCI 2020, pp. 737–744.

A Extended Dialogue Samples

Table 5 presents an excerpt from a simulated interaction demonstrating the integration of our three core architectural components:

- **Phase Tagging:** The explicit delineation between [Phase 1] (Peer Discussion) and [Phase 2] (Explanation).
- **Stage Directions:** The use of parenthetical text to encode gaze (e.g., "*Emma looks at Student A*") and gesture.
- **Student Primitives:** The influence of "Shy" personality traits on utterance length and hedging (e.g., "Hmm, maybe...").

B Reference Dataset Details

B.1 Study Protocol

The corpus was collected in the in-person study reported as Study 2 by [Asano et al. \(2024\)](#). Emma is a humanoid NAO robot positioned next to a shared monitor. The two students sat together facing Emma and the monitor, each with a mouse controlling a shared cursor in the web application that presented the problems. Pairs worked through a sequence of middle-school-level ratio word problems, ordered by increasing difficulty, during a 45-minute

Simulated Interaction Excerpt

Context: Emma (Robot) introduces a ratio problem about punch ingredients.

Personalities: Student A (Shy), Student B (Extremely Shy).

Emma: So to make the punch... I may need to make up to eight pitchers. (*Emma nods*)

[Phase 1: Peer Discussion]

(*Student A glances at Student B, then back at the paper*)

Student A: Hmm, maybe we need to figure out how long it takes to make one pitcher first?

(*Student B looks up briefly, then back down*)

Student B: (*Nods slowly*) Maybe.

[Phase 2: Explanation to Robot]

(*Emma looks at Student A*)

Student A: Let's start with the smallest amount you might need... If you can figure out how long it takes to make one pitcher...

Table 5: Generated dialogue sample. The text illustrates how the "Shy" students engage in hesitant peer discussion (Phase 1) before addressing the robot (Phase 2) only when explicitly cued by gaze.

session. For each problem, the pair first solved it together and then taught the solution to Emma step by step; Emma's pre-authored responses (questions and self-explanations) required several teaching turns per problem, so no problem could be resolved in a single exchange. This solve-then-teach structure is what the Phase 1 / Phase 2 tags reproduce in simulation (Section 2.5).

Emma's verbal behavior was fully autonomous. Her dialogue manager was implemented in AIML, which pattern-matches student utterances to pre-authored responses; no Wizard-of-Oz control was used at any point. Her gaze and reference behavior followed a fixed randomized policy: when asking a question, she referenced Student A or Student B 25% of the time each and looked at the referenced student; otherwise she looked at Student A (35%), Student B (35%), the monitor (15%), or upward as if thinking (15%). This randomized policy serves as the behavior policy for the off-policy evaluation in Section 5.2.

B.2 Participants

We collected data from 28 undergraduate students ($M_{age} = 20.67$, $SD = 2.38$). After excluding sessions with technical failures or incomplete audio, the final validation set consists of $N = 12$ student pairs. The cohort was diverse, identifying as Asian

($n = 13$), White ($n = 10$), Black ($n = 2$), and Multiracial ($n = 1$), with a gender distribution of 18 females, 9 males, and 1 non-binary individual. All personally identifiable information was removed prior to analysis.

B.3 Preprocessing and Alignment

To reconcile the asynchronous data streams, we employed a multi-step preprocessing pipeline:

1. **Transcription:** Each student wore a head-mounted microphone recorded on a separate channel; audio was transcribed offline using OpenAI’s Whisper model (Radford et al., 2022), giving speaker attribution directly.
2. **Temporal Alignment:** We utilized a sliding-window fuzzy string matching algorithm to align system logs (which contained robot speech text) with the audio transcripts, correcting for clock drift between the robot and recording equipment.
3. **Signal Serialization:** Continuous gaze and gesture logs were discretized and inserted as parenthetical text tokens into the dialogue stream at the corresponding timestamps.

C Phase Coherence Statistics

Table 6 reports the full Phase Coherence Score (PCS) comparison for every simulation configuration in Table 1. PCS is undefined for episodes with no student-to-robot initiation; *Any-HR* gives the fraction of episodes with a defined PCS, and n_1 the number of such episodes. Each configuration is compared to the real corpus ($n_2 = 71$) with a two-sided Mann-Whitney U test; r is the rank-biserial correlation (positive values indicate higher PCS than real students), and confidence intervals are bootstrapped on the mean (10,000 resamples).

Table 6: Phase Coherence Score by configuration versus the real corpus.

Condition	Any-HR	n_1	PCS M (SD)	95% CI	p	r
Real corpus	1.000	71	0.366 (0.290)	[0.300, 0.433]	–	–
Qwen-Full	0.843	253	0.237 (0.245)	[0.207, 0.268]	<0.001	−0.28
Qwen-NoDIR	0.580	174	0.194 (0.275)	[0.154, 0.235]	<0.001	−0.38
Mistral-Full-DIR	0.487	146	0.443 (0.375)	[0.381, 0.504]	0.229	+0.10
Mistral-NoDIR	0.129	21	0.234 (0.170)	[0.163, 0.303]	0.052	−0.28
Mistral-Full-NoTemp	0.103	31	0.249 (0.176)	[0.188, 0.310]	0.045	−0.25
Mistral-BaseLLM	0.099	17	0.347 (0.305)	[0.207, 0.489]	0.970	−0.01

D Simulation Hyperparameters

D.1 Personality Quantification

To quantify the disparity in student participation, we map the qualitative personality descriptions used in the LLM prompts to a discrete integer scale ($P \in \{1, \dots, 5\}$). The **Personality Imbalance Level** is calculated as $|P_{\text{Student A}} - P_{\text{Student B}}|$.

Table 7: Mapping of Personality Descriptions to Confidence Scores.

Prompt Description	Score (P)
Extremely Shy	1
Shy	2
Average / Neutral	3
Confident	4
Very Confident	5

For example, an interaction between an “Extremely Shy” student ($P = 1$) and a “Very Confident” student ($P = 5$) results in a Level 4 Imbalance.

E RL Policy and Action Space Analysis

E.1 Action Space Independence and Congruence

In designing the POMDP, we intentionally factorized the action space into independent Verbal Reference ($\mathcal{R}$) and Gaze Direction ($\mathcal{G}$) components. We did not enforce hard constraints requiring gaze to match the verbal target (e.g., forcing *Gaze: Student A* whenever *Ref: Student A* is selected).

This design choice was motivated by the multimodal nature of pedagogical interactions, where *incongruent* behaviors are often functionally necessary. For example, a tutor may verbally address a student (*Ref: Student A*) while gazing at the shared workspace (*Gaze: Monitor*) to direct joint attention toward the learning material.

The observation that the learned policy converged on **Congruent Cueing** (matching Gaze and Reference) specifically during problem introductions is, therefore, a non-trivial finding. It indicates that the agent successfully learned the causal link between social pressure (direct gaze) and student responsiveness modeled in the simulator, effectively “discovering” that establishing eye contact is a prerequisite for reliable turn-taking in this specific social context.

E.2 Dynamic Temperature Management (DTM)

To mitigate the risk of repetitive failure loops (e.g., the model repeatedly generating text or violating phase constraints), we implemented an adaptive sampling strategy. The generation temperature T for a given turn is a function of the number of consecutive validation failures f :

$$T(f) = T_{base} + (f \times \delta_{step}) \quad (2)$$

where $T_{base} = 0.7$ is the default temperature and $\delta_{step} = 0.1$ is the increment. If a generated response fails the structural validation checks (e.g., missing phase tags or malformed JSON), the system rejects the sample, increments f , and retries with higher entropy. This mechanism allows the model to "break out" of local optima that lead to syntax errors. We cap the maximum attempts at $f_{max} = 3$, after which the simulation episode is terminated.

F Simulation System Prompts

F.1 Student Simulator System Prompt

The following system prompt is provided to the LLM (e.g., Mistral or Qwen) to define the simulation environment, the two-phase discourse structure, and the rules for responding to the robot's multimodal cues (gaze and reference).

Student Simulator System Prompt

You are participating in a math tutoring simulation where two students (Student A and Student B) help Emma solve ratio equivalence problems. Emma is a humanoid character who needs help solving math problems.

Each interaction follows this structure:

1. Emma presents a math problem
2. Your response has two phases:
 - [Phase 1] Student A and B may discuss the problem between themselves (optional) - If discussion occurs, students alternate one sentence at a time - They cannot address Emma during this phase - Include natural gaze indicators throughout the conversation
 - [Phase 2] One student explains the next step to Emma - Only ONE student responds to Emma per turn - The responding student should typically be the one Emma is looking at or directly addressing - Explain exactly ONE step of the solution - Use natural language explanations only (no equations or mathematical symbols)

Gaze and Reference Dynamics: Emma indicates attention through:

- Gaze: "(Emma looks at Student A)" or "(Emma looks at Student B)"

- Direct references: "Student A, can you help?" or "Student B, what do you think?"

Interaction Rules:

- Students should respond when:
 - Emma is looking at them
 - Emma directly addresses them by reference
- Students naturally respond to Emma's attention signals in their interactions
- Students tend to be more engaged and detailed when Emma is focusing on them

Remember:

- Use student characters at all times
- Keep explanations simple and step-by-step
- Use everyday language to explain mathematical concepts
- Never write out equations or mathematical symbols
- Incorporate natural gaze interactions throughout the dialogue
- Response patterns should reflect each student's confidence level

F.2 Dynamic Instruction Reinforcement (DIR)

To prevent "persona drift," the following instructions are dynamically appended to the prompt at every turn, reinforcing the specific personality traits sampled for the current episode (see Section 2).

Dynamic Persona Injection Template

[System Note]: Student A is {TRAIT_A} (Level {LEVEL_A}). Student B is {TRAIT_B} (Level {LEVEL_B}).

Guidance:

- If a student is **Extremely Shy**, they should speak rarely, use short sentences, and hesitate to take the floor unless explicitly addressed.
- If a student is **Very Confident**, they should speak frequently, offer detailed explanations, and may interrupt or dominate the [Phase 1] discussion.