Hearing the Unspoken: Simulator-Induced Asymmetric-View Policy Optimization for Proactive Task-Oriented Dialogue

Hongbin Zhang[1] **Ning Gao**[2] **Yuqin Dai**[2] **Ruiyuan Wu**[2] **Jinpeng Wang**[2]
Rena Wei Gao[3] **Bingdong Tan**[4] **Shuzheng Gao**[5] **Chaozheng Wang**[1]*
[1]Vera Praxis [2]Meituan [3]ByteDance
[4]Independent Researcher [5]The Chinese University of Hong Kong

Abstract

Proactive task-oriented dialogue (TOD), such as outbound sales, requires an agent that proactively elicits and addresses the user's concerns within a bounded number of turns, supports an informed voluntary decision, and respects refusal. Yet post-trained LLMs are inherently conservative, and reward-shaping RL (e.g., GRPO) struggles since it only re-weights what an already passive policy samples. We show that, in our evaluated settings, conditioning on the user's latent concerns induces proactive behaviors that the tested sampling budget and outcome-only RL do not reliably induce, establishing these concerns as a substantially stronger training-time signal. To operationalize this finding, we build the **Cognitive User Simulator**, which models each user as a stratified persona comprising observable external traits and hidden internal concerns. The simulator produces faithful and diverse interactions, while emitting per-turn state dynamics that track persuasion progress. We then introduce **Simulator-Induced Asymmetric-View Policy Optimization**, which converts the modeled concerns and the simulation state transition into complementary training objectives: (1) *Asymmetric On-Policy Self-Distillation* that transfers concern-aware behavior from a privileged view of the same policy into its deployable, conversation-only view; and (2) *State-Transition Policy Refinement* where the final decision provides trajectory-level advantage while synchronous state transitions refine turn-level credit direction. Across two real-world-inspired food-delivery settings (*merchant* and *courier* outbound recruitment), our method outperforms strong RL baselines (on Qwen3-8B, concern solving/acceptance rates rise from SEAD's 40.4/36.3% to 52.6/48.5% on merchant and from 36.7/33.5% to 45.6/44.3% on courier), is competitive with the evaluated proprietary LLMs under our evaluation protocol, and generalizes across the tested unseen LLM-based user simulators. Code: `https://github.com/bingozhang-verapraxis/AVPO`.

1 Introduction

Large Language Models (LLMs) have demonstrated strong conversational capabilities, leading interests in applying them to task-oriented dialogue (TOD) [1–4]. Among TOD scenarios, *proactive* settings such as outbound sales are especially challenging compared with conventional reactive ones, where agents respond to users' explicit and direct requests [1]: the agent must initiate or sustain the interaction, guide the dialogue toward a predefined goal, and help users reach an informed decision on a target action [5]. These requirements impose two core demands beyond fluent response generation: agents must **identify users' inherent concerns and implicit needs** and **proactively address them while respecting the user's decision, including refusal**.

*Corresponding author: `wang.chaozheng@verapraxis.ai`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, these demands expose two structural gaps that existing approaches cannot bridge. **Gap1: Current user simulators lack modeling of the implicit concerns.** Existing simulators typically model only surface traits (e.g., tone, politeness) [5–7], which makes simulated users differ mainly in style, causing dialogues to collapse into homogeneous trajectories with limited behavioral diversity. Crucially, such surface traits do not capture the *latent concerns* that actually drive user decisions, making simulator decisions weakly grounded and noisy. **Gap2: LLMs are inherently reactive.** Standard alignment optimizes models to *respond* helpfully rather than to *drive* the conversation, leaving them ill-suited for the initiative-taking required by proactive TOD. Mitigating this gap with current post-training formats proves equally difficult: SFT [8–10] lacks high-quality data, merely mimicking surface utterances; RL methods [5, 7, 11, 12] struggle to credit the specific turns where persuasion occurs, whether relying on sparse outcome rewards or LLM judges that score generic dialogue quality rather than actual progress toward acceptance. These two gaps meet at the same training intersection for LLMs: the user's latent decision process. Without modeling the concerns that guide user decisions, simulators cannot provide grounded decisions, and training methods cannot identify which proactive turns actually resolve user resistance.

To further examine whether this missing variable limits current LLM policies, we conduct a pilot study to evaluate the capability and proactivity of LLMs in proactive TOD. Figure 1 summarizes the results in the (Proactivity, Acceptance Rate) plane. Vanilla models (e.g., Qwen3-4B/8B) anchor at low operating points, and test-time scaling with Pass@16 leaves them clustered nearby, suggesting that the tested sampling budget does not elicit the required proactive behaviors. Reward-shaping RL (e.g., GRPO) consequently exhibits a similar ceiling, as it can only re-weight behaviors already sampled by the policy. By contrast, conditioning the same agents on users' latent concerns yields substantial gains in both proactivity and acceptance (dashed arrows). In our evaluated settings, the tested sampling budget and outcome-only RL therefore did not reliably induce the same proactive behaviors, whereas latent-concern access provided a substantially stronger training signal for learning proactive policies.

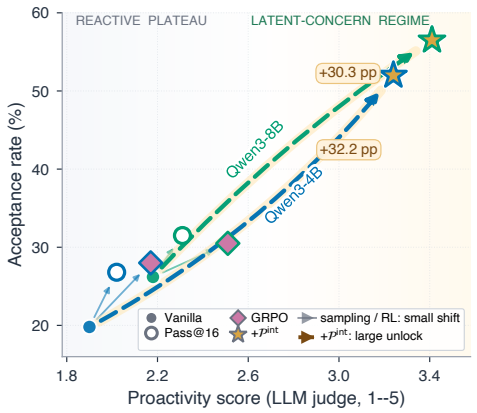

Figure 1: **Pilot study: latent concerns unlock proactive persuasion.** Starting from the same Vanilla baseline, sampling (thin solid) yields only a marginal Pass@16 and GRPO shift, while conditioning on users' latent concerns $\mathcal{P}^{\text{int}}$ (thick dashed) propels both models into the high-proactivity / high-acceptance regime.

Guided by this finding, we make two linked but separable contributions: CUS supplies a concern-grounded training environment and state-transition signals, and SI-AVPO converts its privileged concern view and these signals into a deployable, dialogue-only policy. The **Cognitive User Simulator (CUS)** models each user as a stratified persona comprising observable external traits, hidden internal concerns $\mathcal{P}^{\text{int}}$, and synchronous state dynamics. By explicitly modeling the latent concerns that anchor user decisions, CUS produces behaviorally diverse and faithful interactions, providing a high-fidelity training environment that surface-trait-only simulators cannot offer. It further emits simulation feedback—the final accept/reject outcome and per-turn internal state transitions—that records how user receptiveness changes in each turn. We then introduce **Simulator-Induced Asymmetric-View Policy Optimization** to convert the modeled concerns and the simulation state into complementary training objectives. (1) *Asymmetric On-Policy Self-Distillation (AOPD)* evaluates a *single* policy under two asymmetric information views: a privileged view that additionally conditions on $\mathcal{P}^{\text{int}}$, and a deployable view that observes only the dialogue history. The privileged view's concern-aware behavior is progressively distilled into the deployable one, requiring no separate teacher model and enabling the agent to implicitly infer and address latent concerns at inference time. (2) *State-Transition Policy Refinement (STPR)* complements this with evaluative policy-gradient signals: the final decision establishes trajectory-level advantage while per-turn state transitions refine step-level credit direction—all derived from CUS without external reward models or judges.

Extensive experiments validate both components. Human evaluation confirms that CUS produces substantially more diverse and truthful dialogue topics and more grounded decisions than prior LLM

simulators. Built on CUS, our trained policy achieves higher task success than RL baselines in two real-world-inspired outbound-call settings and is competitive with the evaluated proprietary LLMs under our evaluation protocol, as rated by LLM judges and human evaluators, and a blinded live expert role-play shows consistent trends beyond simulator-internal metrics.

We make the following contributions:

- We propose Cognitive User Simulator (CUS), which models each user as a stratified persona with hidden internal concerns $\mathcal{P}^{\text{int}}$ and synchronous state dynamics, calibrated on 100K+ real-world outbound-call logs, producing faithful and behaviorally diverse interactions.
- We introduce Simulator-Induced Asymmetric-View Policy Optimization, which converts the modeled concerns and simulation state from CUS into complementary training objectives via Asymmetric On-Policy Self-Distillation and State-Transition Policy Refinement, with no external reward model, judge, or another teacher model.
- Experiments in two real-world-inspired food-delivery proactive-TOD settings (*merchant* and *courier* outbound recruitment) show that our method outperforms strong RL baselines and is competitive with the evaluated proprietary LLMs (e.g., Claude-Sonnet-4.5, GLM-5.1) under our evaluation protocol, while remaining robust across the tested unseen LLM-based user simulators.

2 Related Work

Task-oriented dialogue (TOD) has evolved from sequence-to-sequence architectures [13, 14] to LLM-driven paradigms that leverage in-context reasoning and instruction following [15, 16], with reinforcement learning increasingly applied for policy optimization in multi-turn settings [7, 17, 18]. On the post-training side, alignment has progressed from RLHF/DPO [19, 20] through value-free alternatives like GRPO [11, 12], multi-turn RL [21, 22], and distillation-based methods such as OPD [23] and OPSD [24]. Preference-elicitation work teaches LMs to ask informative or clarifying questions and to infer user preferences at interaction time [25–27]; we instead use a structured latent-concern model as privileged training-time information for simulator-grounded policy learning. However, existing TOD methods lack proactive, multi-dimensional user modeling, and current post-training recipes struggle to internalize latent user state for persuasive policy learning. We address both gaps with our framework; Extended discussions are in Appx. A.

3 Methodology

Motivated by the finding that latent concerns are a key missing signal for proactive TOD (§1), we build a framework that models these concerns inside the simulator and converts them into training signals as shown in Fig. 2: (1) The **Cognitive User Simulator (CUS)** (§3.2) represents each user with observable external traits, hidden internal concerns, and concern-driven state dynamics, yielding both a faithful training environment and two categories of natural byproduct: the privileged internal persona $\mathcal{P}^{\text{int}}$ and simulation feedback (final decision and per-turn state transitions). (2) **Simulator-Induced Asymmetric-View Policy Optimization (SI-AVPO)** (§3.3) then converts these signals into complementary training objectives: *Asymmetric On-Policy Self-Distillation (AOPD)* evaluates the same policy under a privileged view that additionally conditions on $\mathcal{P}^{\text{int}}$ and progressively distills the resulting concern-aware behavior into the deployable, dialogue-only view; *State-Transition Policy Refinement (STPR)* complements AOPD with policy-gradient credit, where the final decision establishes trajectory-level advantage while per-turn state transitions refine step-level credit direction—all derived from CUS without any external reward model or judge.

3.1 Problem Definition

We formalize proactive task-oriented dialogue as an episodic, undiscounted, turn-level Markov Decision Process $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, R\}$. Each episode samples a user profile $\mathcal{P}$ that parameterizes the transition dynamics; the detailed design are described in §3.2. The profile governs all user-side dialogue behavior and decisions but is never exposed to the agent.

States and actions. The state space $\mathcal{S}$ consists of variable-length dialogue histories. The initial state $s_0 = [x_0]$ contains only the system instruction x_0 (task description); the agent produces the opening utterance $y_0 \sim \pi_\theta(\cdot \mid s_0)$ *before* any user response. At subsequent turns, the state $s_k = [x_0, y_0, u_1, y_1, \ldots, u_{k-1}, y_{k-1}, u_k]$ concatenates the full history up to the user's k-th utterance u_k, and the agent responds with $y_k \sim \pi_\theta(\cdot \mid s_k)$. The action space $\mathcal{A} = \mathcal{V}^*$ comprises sequences

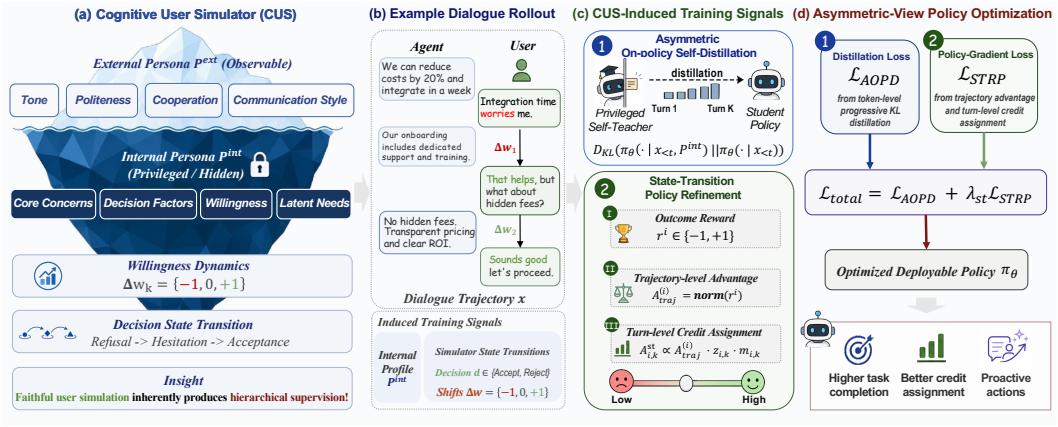

Figure 2: Overview of our framework.

over the token vocabulary. Given s_k and agent response y_k, the user simulator generates u_{k+1} conditioned on $\mathcal{P}$ and the history, yielding $s_{k+1} = [s_k,\ y_k,\ u_{k+1}]$; the resulting transition is denoted $P_{\mathcal{P}}(s_{k+1} \mid s_k, y_k)$.

Reward and objective. The episode runs for at most K turns and terminates with a binary user decision $d \in \{\text{ACCEPT}, \text{REJECT}\}$. A terminal reward $r \in \{+1, -1\}$ is assigned for acceptance and rejection, respectively; the agent's objective is to maximize the expected reward:

$$\max_{\pi_\theta} \ \mathbb{E}_{\mathcal{P},\ \pi_\theta}\big[r\big]. \tag{1}$$

3.2 Cognitive User Simulator

To provide a faithful environment that produces behaviorally diverse and grounded interactions, we propose a stratified user simulator framework that models the latent concerns that anchor user decisions and couples state dynamics to concern progress.

Three-Layer Persona Framework. We instantiate CUS for proactive TOD tasks in collaboration with the operations team of a major food-delivery platform; the persona schema and the concern bank described below are co-designed with domain specialists from aggregate patterns in 100K+ outbound-call records, which also inform the simulator's rules for language, pacing, and reactions. The records are not used as a training corpus or released (Appendix M). Each user is represented as a stratified profile $\mathcal{P} = \big(\mathcal{P}^{\text{bg}},\ \mathcal{P}^{\text{ext}},\ \mathcal{P}^{\text{int}}\big)$ whose three layers play distinct functional roles. The (1) *background* $\mathcal{P}^{\text{bg}}$ specifies the situational frame at call onset (e.g., demographic attributes) and is naturally exposed through the call. The (2) *external persona* $\mathcal{P}^{\text{ext}}$ governs surface behavior along five observable axes (time pressure, courtesy, communication style, cooperation tendency, technical familiarity) and shapes the texture of every user utterance the agent receives.

The (3) *internal persona* $\mathcal{P}^{\text{int}}$ encodes the user's latent concerns across financial, psychological, and operational facets (full taxonomy in Appendix N.2). Each $\mathcal{P}^{\text{int}}$ dimension is grounded by a curated *concern bank*: every concern is annotated with unlock conditions that specify when it can be resolved, and anti-patterns that penalize harmful agent behavior. $\mathcal{P}^{\text{int}}$ directly governs both the per-turn state dynamics and the terminal decision: willingness can only rise meaningfully through concern resolution, and acceptance is gated by sufficient concern coverage—ensuring that the simulator's decisions reflect genuine concern progress rather than free-form LLM judgments. The full concern schema and state-transition logic are described in Appendix N.1.

Simulation State Transitions. Beyond the static persona, CUS maintains a scalar *willingness* state that tracks how close the user is to acceptance and is updated turn by turn. At each agent turn k, the simulator emits a textual user response u_k, a discrete action $a_k \in \{\text{CHAT}, \text{HANGUP}\}$, a signed willingness shift Δw_k, and per-concern state transitions $\Delta \mathcal{C}_k$ that track each concern through monotone states (UNRESOLVED $\rightarrow$ PARTIALLY-ADDRESSED $\rightarrow$ RESOLVED). Importantly, Δw_k is not a free-form LLM score: it is rule-bounded by concern progression, so positive shifts require actual concern resolution and anti-pattern hits induce penalties. The episode terminates with a binary decision $d \in \{\text{ACCEPT}, \text{REJECT}\}$, gated by willingness and concern coverage.

Summary. This design yields one privileged state signal and two natural state-transition signals, which are all available without any external annotation:

$$\Big(\underbrace{\mathcal{P}^{\text{int}}}_{\text{privileged internal state}} , \underbrace{d, \{\Delta w_k\}_{k=1}^{K}}_{\text{simulation state-transitions}} \Big). \tag{2}$$

The internal state, which specifies the latent concerns invisible to the deployed agent, supports asymmetric self-distillation; the final decision and per-turn willingness shifts provide trajectory- and turn-level credit assignment. The next section composes these signals into the training objective.

3.3 Simulator-Induced Asymmetric-View Policy Optimization

To utilize the user's invisible latent decision process induced by CUS and derive training signals, we propose **Simulator-Induced Asymmetric-View Policy Optimization (SI-AVPO)**.

Overall Objective. SI-AVPO has two complementary components that convert the aforementioned signals into training objectives, respectively. First, *Asymmetric On-Policy Self-Distillation (AOPD)* transfers concern-aware behavior from a privileged view of the same policy into its deployable dialogue-only view. Second, *State-Transition Policy Refinement (STPR)* uses the simulator's final decision and synchronous state (i.e., willingness) transitions to refine on-policy credit assignment:

$$\mathcal{L}_{\text{SI-AVPO}} = \mathcal{L}_{\text{AOPD}}(\mathcal{P}^{\text{int}}) + \lambda_{\text{st}}\, \mathcal{L}_{\text{STPR}}(d, \{\Delta w_k\}_{k=1}^{K}). \tag{3}$$

Asymmetric On-Policy Self-Distillation. AOPD aims to turn the internal persona exposed by CUS at training time into token-level supervision for a dialogue-only policy. We define two information views of the same policy π_θ: a *privileged teacher* view whose context is augmented with the user's internal persona $\mathcal{P}_i^{\text{int}}$, and a *deployable student* view that observes only the task instruction and dialogue history. For each student-generated token in an on-policy rollout, we compute the probability distributions under both views and minimize their divergence, thereby distilling concern-aware behavior from the teacher into the student:

$$\mathcal{L}_{\text{AOPD}} = \frac{1}{N} \sum_{i,t} \phi(k(t), K) D_{\text{KL}}\big(\text{sg}\big[\pi_\theta\big(\cdot \mid x_{i,t}, \mathcal{P}_i^{\text{int}}\big)\big] \;\big\|\; \pi_\theta(\cdot \mid x_{i,t})\big)\,. \tag{4}$$

Here, N is the number of assistant tokens, K is the maximum turns, $k(t)$ maps token t to its dialogue turn, and $\phi(k(t), K)$ is a nonnegative turn-dependent distillation weight. The stop-gradient operator $\text{sg}[\cdot]$ freezes the teacher distribution as a fixed target, so gradients flow only through the student view. Because both views share the same π_θ, no separate teacher model is needed; at inference time the policy operates under the student view alone, without any dependence on $\mathcal{P}^{\text{int}}$.

State-Transition Policy Refinement. While AOPD provides directive supervision that teaches the student *what* a concern-aware agent would say, it leaves the question of *whether* a particular attempt actually steers the user's acceptance, e.g., a turn may probe the right concern at the wrong moment that triggers user resistance. Furthermore, since $\mathcal{P}^{\text{int}}$ is unavailable at inference, perfect imitation of the privileged teacher is inherently unattainable; the student must explore and refine its own dialogue strategies under the constrained information view, yet distillation alone offers no evaluative signal to guide this exploration.

Therefore, we propose STPR to address both limitations by introducing evaluative credit grounded in the simulator's state dynamics. STPR uses CUS state transition dynamics to assign trajectory credit. Formally, the final decision defines a normalized trajectory advantage $A_{\text{traj}}^{(i)} = \frac{r_i - \mu_r}{\sigma_r + \varepsilon_{\text{norm}}}$, setting the trajectory-level update direction and scale. Here, r_i is the terminal reward for rollout i, μ_r and σ_r are the mean and standard deviation of rewards within the rollout group, and $\varepsilon_{\text{norm}}$ is a small positive constant for numerical stability. The per-turn willingness shift $\Delta w_{i,k}$ then adjusts this signal: a transition that directionally aligns with the final outcome reinforces the trajectory direction, a contradictory transition flips it, and a zero transition leaves the trajectory credit unchanged. Let $\Delta w_{\max}$ denote the maximum absolute willingness shift over all turns in the rollout group. For turns with a nonzero willingness shift, let $m_{i,k} = |\Delta w_{i,k}| / \Delta w_{\max}$ denote the magnitude gate and $z_{i,k} = \text{sgn}(A_{\text{traj}}^{(i)} \Delta w_{i,k})$ denote the alignment sign. The state-transition advantage is:

$$A_{i,k}^{\text{st}} = A_{\text{traj}}^{(i)} z_{i,k} m_{i,k} \exp(\tau z_{i,k})\,, \quad \Delta w_{i,k} \neq 0. \tag{5}$$

The magnitude gate $m_{i,k}$ measures how much the simulator state changes at turn k, and the temperature τ controls the strength of outcome–transition polarization: larger values amplify turns whose

Algorithm 1: Simulator-Induced Asymmetric-View Policy Optimization (SI-AVPO).

Input : policy π_θ; CUS persona sampler $\mathcal{P}$; group size G; max turns K; inner epochs E
Hyper : $\varepsilon_{\text{clip}}, \varepsilon_{\text{norm}}, \phi(k, K), \tau, \lambda_{\text{st}}$
Output : deployable policy π_θ

for each training iteration **do**
  $\theta_{\text{old}} \leftarrow \theta$
  `// Phase 1: collect CUS rollouts under the deployable view`
  Sample $\mathcal{P}_i \sim \mathcal{P}$
  for $i = 1$ **to** G **do**
    Run a K-turn dialogue with $\pi_{\theta_{\text{old}}}(\cdot \mid x)$
    Record $\{(x_{i,t}, y_{i,t}, k(t))\}$, hidden profile $\mathcal{P}_i^{\text{int}}$, shifts $\{\Delta w_{i,k}\}_{k=1}^{K}$, and reward r_i
  end
  `// Phase 2: compute simulator-grounded STPR credit`
  $A_{\text{traj}}^{(i)} \leftarrow (r_i - \mu_r)/(\sigma_r + \varepsilon_{\text{norm}})\ \forall i, \quad \Delta w_{\max} \leftarrow \max_{j\in[G],\, \ell\in[K]} |\Delta w_{j,\ell}|$
  foreach *turn* (i, k) **do**
    $z_{i,k} \leftarrow \text{sgn}(A_{\text{traj}}^{(i)} \Delta w_{i,k}), \quad m_{i,k} \leftarrow |\Delta w_{i,k}|/\Delta w_{\max}$
    $A_{i,k}^{\text{st}} \leftarrow \begin{cases} A_{\text{traj}}^{(i)}, & \Delta w_{i,k} = 0, \\ A_{\text{traj}}^{(i)} z_{i,k} m_{i,k} \exp(\tau z_{i,k}), & \Delta w_{i,k} \neq 0; \end{cases}$
    Broadcast $A_{i,k}^{\text{st}}$ to all assistant tokens in turn k
  end
  `// Phases 3-4: privileged AOPD loss and joint update`
  for $e = 1$ **to** E and each minibatch $\mathcal{B}$ **do**
    foreach $(i, t) \in \mathcal{B}$ **do**
      $q_{i,t}^{\text{priv}} \leftarrow \text{sg}\big[\pi_\theta(\cdot \mid x_{i,t}, \mathcal{P}_i^{\text{int}})\big], \quad p_{i,t} \leftarrow \pi_\theta(\cdot \mid x_{i,t}), \quad \rho_{i,t} \leftarrow \dfrac{\pi_\theta(y_{i,t} \mid x_{i,t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x_{i,t})}$
    end
    $\mathcal{L}_{\text{AOPD}}^{\mathcal{B}} \leftarrow \dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{(i,t)\in\mathcal{B}} \phi(k(t), K) D_{\text{KL}}\Big(q_{i,t}^{\text{priv}} \Big\| p_{i,t}\Big)$
    $\mathcal{L}_{\text{STPR}}^{\mathcal{B}} \leftarrow -\dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{(i,t)\in\mathcal{B}} \min\Big(\rho_{i,t} A_{i,k(t)}^{\text{st}}, \text{clip}(\rho_{i,t}, 1-\varepsilon_{\text{clip}}, 1+\varepsilon_{\text{clip}}) A_{i,k(t)}^{\text{st}}\Big)$
    $\theta \leftarrow \theta - \nabla_\theta \Big(\mathcal{L}_{\text{AOPD}}^{\mathcal{B}} + \lambda_{\text{st}} \mathcal{L}_{\text{STPR}}^{\mathcal{B}}\Big)$
  end
end
return π_θ

willingness shift aligns with the final outcome and more strongly suppress contradictory turns. The alignment sign $z_{i,k}$ determines the credit direction: it is positive when the local willingness shift agrees with the terminal trajectory advantage and negative when it contradicts it. The exponential factor then further amplifies aligned transitions and suppresses contradictory ones. If $\Delta w_{i,k} = 0$, we directly set $A_{i,k}^{\text{st}} = A_{\text{traj}}^{(i)}$: the simulator provides no local evidence for changing the credit direction, so the turn keeps the original trajectory credit. STPR therefore anchors credit to the trajectory-level advantage while assigning aligned turns same-signed credit and contradictory turns opposite-signed credit according to the simulator's state dynamics. It relies mainly on the *direction* of each willingness shift; the magnitude gate and the exponential factor provide smaller refinements (§5.2). Broadcasting $A_{i,k}^{\text{st}}$ to all assistant tokens in turn k gives the clipped refinement loss

$$\mathcal{L}_{\text{STPR}} = -\frac{1}{N} \sum_{i,t} \min\Big(\rho_{i,t}\, A_{i,k(t)}^{\text{st}},\ \text{clip}(\rho_{i,t}, 1-\varepsilon_{\text{clip}}, 1+\varepsilon_{\text{clip}})\, A_{i,k(t)}^{\text{st}}\Big), \tag{6}$$

where $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x_{i,t})/\pi_{\theta_{\text{old}}}(y_{i,t} \mid x_{i,t})$. Algorithm 1 summarizes the complete training loop.

4 Experimental Setup

Baselines. We compare SI-AVPO against two categories of baselines. **(1) Large Models.** We select five state-of-the-art proprietary LLMs—GPT-4.1 [28], DeepSeek-V3.2 [29], GLM-5.1 [30], Kimi-K2.5 [31], and Claude-Sonnet-4.5 [32]—with the same task instructions as strong references. **(2) RL-based Methods.** Starting from two foundation models, Qwen3-4B and Qwen3-8B [33], we apply several representative reinforcement learning algorithms: GRPO [12], PPO [34], DAPO [11], and

Table 1: Comparison of model performance across different scenes on *FoodDeliverService*, reported as mean ± standard deviation over three independent evaluation runs. CSR: Concern Solving Rate (%); Acc: Acceptance Rate (%); Comm.: Communication Quality; Logic: Logic Quality; Proact.: Proactive Capability. **Bold**: best within each group; underline: second best within each group.

Method	Scene 1 (Merchant)					Scene 2 (Courier)				
	Task Score		Dialogue Metric			Task Score		Dialogue Metric		
	CSR (%)↑	Acc (%)↑	Comm.↑	Logic↑	Proact.↑	CSR (%)↑	Acc (%)↑	Comm.↑	Logic↑	Proact.↑
Large Models										
GPT-4.1	36.1±2.7	34.7±2.2	2.93±0.15	3.02±0.18	2.12±0.30	30.7±2.9	22.8±1.7	2.66±0.21	2.91±0.19	1.79±0.08
DeepSeek-v3.2	45.7±1.8	39.5±2.4	3.04±0.30	3.13±0.17	2.45±0.15	32.6±2.1	28.0±1.9	2.54±0.16	3.27±0.17	1.96±0.12
GLM-5.1	46.2±2.9	42.3±1.8	3.32±0.13	3.44±0.20	2.78±0.22	38.7±2.2	38.7±2.2	2.86±0.15	3.56±0.15	2.12±0.26
Kimi-K2.5	49.0±2.4	39.8±2.4	3.26±0.13	3.51±0.24	2.67±0.26	39.0±2.7	38.8±2.1	2.79±0.16	3.45±0.15	2.21±0.16
Claude-Sonnet-4.5	**51.1**±2.4	**50.8**±2.5	**3.56**±0.16	**3.76**±0.22	**3.09**±0.13	**45.0**±1.7	**43.2**±2.1	**3.03**±0.09	**3.75**±0.16	**2.67**±0.15
Qwen-3-4B										
Base	22.5±1.9	20.2±2.5	2.23±0.19	2.15±0.18	1.90±0.21	20.8±3.0	18.5±1.2	1.96±0.21	2.34±0.19	1.76±0.22
+GRPO	28.4±2.9	28.0±2.8	2.47±0.30	2.61±0.11	2.17±0.11	24.3±1.8	22.3±1.4	2.13±0.17	2.55±0.14	1.86±0.30
+PPO	26.5±1.9	22.3±2.0	2.36±0.24	2.59±0.12	2.22±0.09	22.7±2.3	19.8±2.1	2.06±0.23	2.44±0.22	1.75±0.19
+DAPO	29.7±1.7	29.5±2.5	2.60±0.18	2.74±0.13	2.33±0.30	23.8±2.3	26.3±1.6	2.25±0.24	2.81±0.26	2.10±0.20
+SEAD	33.7±2.0	32.7±1.7	2.95±0.18	3.22±0.21	2.58±0.18	30.2±3.3	32.5±2.5	2.52±0.12	3.21±0.14	2.26±0.22
+SI-AVPO	**51.9**±1.7	**46.3**±2.4	**3.25**±0.10	**3.56**±0.07	**3.11**±0.11	**44.6**±2.3	**42.0**±2.1	**2.87**±0.10	**3.52**±0.11	**2.63**±0.12
Qwen-3-8B										
Base	30.6±3.3	26.2±4.5	2.66±0.17	2.48±0.25	2.18±0.30	23.5±1.2	22.2±2.0	2.35±0.30	2.51±0.15	2.00±0.24
+GRPO	33.2±1.5	30.5±2.2	2.69±0.12	2.88±0.18	2.51±0.21	29.6±2.6	24.7±1.9	2.41±0.14	3.03±0.13	2.11±0.25
+PPO	31.8±4.1	26.8±2.8	2.60±0.27	2.81±0.13	2.26±0.20	25.2±2.5	24.8±0.8	2.40±0.12	2.87±0.30	2.13±0.09
+DAPO	36.3±3.4	34.3±1.7	2.81±0.11	3.37±0.22	2.41±0.19	30.7±2.0	28.7±3.4	2.59±0.12	3.19±0.19	2.28±0.22
+SEAD	40.4±2.3	36.3±1.8	3.23±0.14	3.56±0.25	2.63±0.25	36.7±2.6	33.5±1.9	2.75±0.15	3.51±0.18	2.45±0.19
+SI-AVPO	**52.6**±2.2	**48.5**±2.0	**3.61**±0.25	**3.71**±0.10	**3.18**±0.16	**45.6**±1.7	**44.3**±1.6	**3.06**±0.11	**3.66**±0.12	**2.94**±0.11

SEAD [5]. All RL baselines are trained with the same user simulator and environment configuration for a fair comparison.

Datasets. We instantiate SI-AVPO on two real-world-inspired proactive outbound-call tasks from our business: *Merchant Promotion* (recruiting new merchants into a paid traffic campaign) and *Courier Regional Bonus* (enrolling couriers into a region-specific rush-hour bonus program). For each task we sample $10{,}000$ personas for training and a disjoint set of 200 personas for evaluation, so test-time performance reflects generalization to unseen profile combinations rather than memorization. During persona construction, a value is sampled for each internal-profile dimension, and the user's specific concerns are then sampled uniformly at random from the concern-bank entries of the sampled values, with the number of concerns drawn from $\mathrm{Uniform}\{3, 6\}$, yielding diverse difficulty levels across episodes. The full persona schema, dimension breakdown, and concern-pool construction are detailed in Appendix N.2.

Evaluation Metrics. We measure two simulator-derived signals from the simulator's internal state transitions: the **Acceptance Rate (Acc)**, the percentage of dialogues ending in user acceptance, and the **Concern Solving Rate (CSR)**, the proportion of the user's hidden concerns successfully addressed during the interaction. To assess interaction quality beyond task completion, we employ three strong third-party LLMs (Gemini-3.1-Pro [35], GPT-5.4 [36], and Claude-Opus-4.6 [37]) as judges to score each dialogue along three dimensions: **Communication**, **Logic**, and **Proactivity**. Full metric definitions and judge protocols are provided in Appendix C.

Implementation Details. All experiments are implemented using veRL [38]. For RL training, we use 8× NVIDIA A100 (80GB) GPUs. The user simulator is served on 16× NVIDIA H20 GPUs to support high-throughput parallel rollout generation. We use Qwen3.5-115B [39] as the default user simulator during training and Claude-Sonnet-4.5 for evaluation. For evaluation, each trained policy is evaluated in *three* independent runs. For LLM-as-judge metrics, each dialogue is scored by the three judges introduced above independently, and we report the average scores. Details are provided in Appx. D, and the statistical procedure in Appx. I.

5 Experiment Results

5.1 Main Results

We first evaluate SI-AVPO against two groups of baselines: representative RL algorithms built on the same Qwen backbones and flagship proprietary LLMs prompted with the same task instructions. Table 1 summarizes both task success and dialogue-quality metrics on the merchant and courier outbound-call scenes.

Among the RL baselines, general-purpose algorithms (GRPO, DAPO, PPO) yield modest improvements over the base model in task scores and dialogue quality, yet their gains on the proactivity remain limited. For example, on Qwen3-8B, DAPO raises Proact. from 2.18 to only 2.41 on the merchant scene. This pattern suggests that outcome-driven RL can refine surface-level communication skills and help the agent better address user concerns once they are raised, but it struggles to induce genuinely proactive behaviors such as anticipating latent objections or steering the conversation toward unspoken needs. SEAD advances further through its curriculum-based self-evolution, achieving stronger performance. Nevertheless, SI-AVPO consistently and substantially outperforms all RL baselines across both scenes and both backbone sizes. On Qwen3-8B, SI-AVPO significantly surpasses the best RL baseline (SEAD) by +12.2% CSR, +12.2% Acc, and +0.55 Proact. on the merchant scene, and by +8.9% CSR, +10.8% Acc, and +0.49 Proact. on the courier scene (statistical tests in Appx. I). The improvement in proactivity indicates that our SI-AVPO provides direct supervision for initiative-taking behaviors that conventional RL rarely discovers on its own: actively probing hidden concerns, adapting to user hesitation, and steering conversations toward decisions rather than passively responding.

Notably, SI-AVPO outperforms four of the five flagship LLMs, including GPT-4.1, DeepSeek-v3.2, GLM-5.1, and Kimi-K2.5, on most metrics across both scenes, even with the smaller Qwen3-4B backbone. With Qwen3-8B, SI-AVPO further rivals the strongest flagship and proprietary model, Claude-Sonnet-4.5, surpassing it on CSR (52.6% vs. 51.1%), Proact. (3.18 vs. 3.09) on the merchant scene, and on both CSR and Acc on the courier scene, while trailing only marginally on Acc and Logic for merchants. Since flagship models of this scale are impractical for real-time outbound calling due to latency, cost, and data-governance constraints, the key takeaway is that a deployable-size policy trained with SI-AVPO is competitive with the evaluated, much larger proprietary LLMs on the two simulated food-delivery tasks under our evaluation protocol.

5.2 Ablation Study

Table 2 presents the ablation results, from which we draw the following observations.

Table 2: Ablation on *Merchant* (Qwen3-4B). STPR only: GRPO with terminal and Δw feedback; GRPO: terminal outcome only. –: not evaluated.

Variant	CSR(%)	Acc(%)	Comm.	Logic	Proact.
SI-AVPO (Full)	**51.9**	**46.3**	**3.25**	**3.56**	3.11
Removing AOPD components					
w/o AOPD (→STPR only)	35.4 ↓16.5	34.5 ↓11.8	2.81	3.01	2.35
w/o AOPD + Turn (→GRPO)	28.4 ↓23.5	28.0 ↓18.3	2.47	2.61	2.17
Removing STPR components					
w/o Turn-Level Reward	46.7 ↓5.2	42.2 ↓4.1	3.01	3.21	3.14
w/o STPR (→AOPD)	44.6 ↓7.3	38.7 ↓7.6	2.85	2.98	**3.23**
w/o STPR + Asym (→OPD)	37.6 ↓14.3	32.8 ↓13.5	2.66	3.01	2.39
Matched-feedback controls (no privileged concerns)					
PPO (terminal + Δw)	38.2 ↓13.7	36.3 ↓10.0	–	–	–
SFT (2K successes)	32.7 ↓19.2	31.2 ↓15.1	–	–	–
DPO (2K pairs)	34.3 ↓17.6	33.8 ↓12.5	–	–	–

AOPD is essential for effective exploration. Removing AOPD while retaining STPR ("w/o AOPD") causes CSR to drop by 16.5% and Acc by 11.8%. Crucially, the steepest single-metric decline is on Proact. (-0.76). This decrease confirms that AOPD provides a *directional* training signal: by distilling the privileged teacher's targeted strategies into the student, it specifically guides the policy toward initiative-taking behaviors rather than merely improving surface fluency. Further stripping the turn-level reward ("w/o AOPD + Turn-Level", equivalent to vanilla GRPO) yields an additional 7.0% CSR and 6.5% Acc decline, indicating that sparse outcome rewards alone are insufficient to discover effective proactive strategies.

Matched-feedback controls. Controls with the same CUS feedback but no privileged concerns include PPO with terminal and Δw feedback, SFT on 2K successful trajectories, and DPO on 2K matched-context pairs (trajectories from the default user-simulator model). The strongest, PPO, trails SI-AVPO by 13.7 CSR and 10.0 Acc points. As STPR only and PPO get the same feedback as SI-AVPO without privileged concerns, their gap to SI-AVPO reflects privileged-view distillation rather than a richer reward; SFT and DPO show that replaying successful trajectories does not suffice.

STPR grounds AOPD's proactive drive in context. Removing the turn-level reward ("w/o Turn-Level Reward") lowers CSR by 5.2% and Acc by 4.1%, with noticeable drops in Comm. (-0.24) and Logic (-0.35). Further removing the entire STPR component ("w/o STPR", i.e. AOPD only) widens the gap to -7.3% CSR and -7.6% Acc, while Comm. and Logic continue to deteriorate

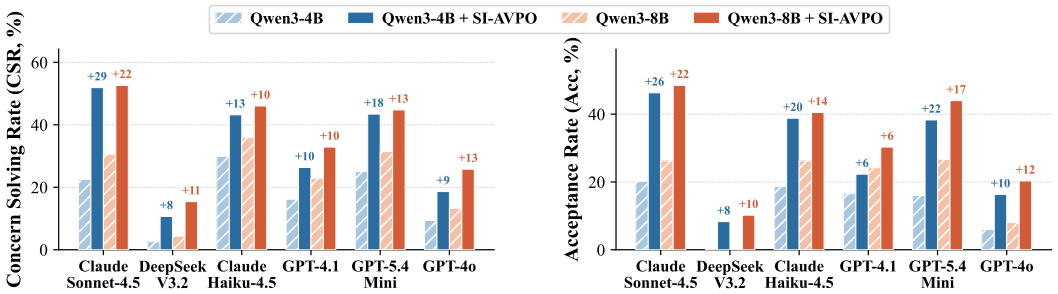

Figure 3: Generalization across different user simulators on *Merchant* scene.

(-0.40 and -0.58, respectively). Interestingly, the Proact. score *increases* in both variants (3.14 and 3.23 vs. 3.11 in the full model), revealing that AOPD alone drives the policy toward an *overly aggressive* persuasion style: it eagerly pushes proposals and steers toward acceptance, but neglects the user's intermediate reactions and conversational context. STPR corrects this tendency by supplying fine-grained, per-turn evaluative feedback that grounds each proactive move in the dialogue state, encouraging the agent to balance initiative with attentive, context-aware responses.

The role of information asymmetry. To isolate whether the gain of AOPD stems from *privileged information* or merely from the distillation mechanism itself, we replace AOPD with a vanilla OPD variant ("OPD only"): the teacher is upgraded to a stronger Qwen3-32B model but observes only the same dialogue history as the student, without access to the hidden internal persona $\mathcal{P}^{\text{int}}$. Despite its substantially larger capacity, OPD yields 7.0% lower CSR and 5.9% lower Acc than AOPD, with Proact. dropping sharply from 3.23 to 2.39 (-0.84). This confirms that the critical ingredient is not teacher capacity but *information asymmetry*—access to the user's latent concerns enables the teacher to demonstrate targeted, concern-aware strategies that a stronger but equally uninformed teacher cannot discover from dialogue alone.

Reliability of the willingness signal. Re-running the simulator three times on 300 fixed (persona, prefix, turn) tuples gives the same sign of Δw in 98.3% of tuples (mean normalized SD 0.12), and blinded experts' direction labels agree with $\text{sgn}(\Delta w)$ in 92% of cases. In the STPR-only setting (35.4/34.5 CSR/Acc), permuting (28.4/25.1) or noising (30.1/29.6) the shifts degrades results substantially, whereas sign-only shifts, removing $\exp(\tau z)$, or zero credit for $\Delta w = 0$ costs at most 1.2 points; STPR thus relies mainly on correct directions. One-factor sweeps of ϕ, τ, and λ_{st} keep CSR within 46.7–51.9%, with defaults best (Appx. H).

5.3 Generalization on Different User Simulator

Figure 3 evaluates the transferability of the learned dialogue policy to user simulators unseen during training. An important observation is that different LLMs, even when instantiated with identical persona profiles and instructions, exhibit markedly different levels of strictness in role-playing the user. For instance, DeepSeek-V3.2 proves to be an exceptionally strict simulator—both Qwen3-4B and Qwen3-8B achieve near-zero Acceptance Rate against it in the base setting. This variance reflects genuine differences in how each LLM interprets and enacts the user persona rather than differences in task difficulty, as all simulators share the same scenario configuration. To rigorously quantify these behavioral differences, we conduct a per-simulator quality assessment in Appendix E, combining human evaluation with LLM-as-judge assessment to measure each simulator's adherence to the prescribed persona and its decision Rationality.

Despite this substantial variation in simulator strictness, SI-AVPO yields consistent improvements across *every* tested simulator. On average, SI-AVPO improves CSR by 14.8% and Acc by 15.5% on Qwen3-4B, and by 13.2% and 13.8% on Qwen3-8B, with gains observed on all six simulators without exception. This uniformity across diverse model families, parameter scales, and conversational styles indicates that SI-AVPO does not overfit to the idiosyncrasies of a single training-time simulator. As all tested simulators retain the CUS persona and concern specification, this shows robustness within CUS-like environments (§5.5 tests human role-players).

Table 3: User simulator fidelity on 500 dialogues per simulator (*Merchant*) with Qwen3.5 115B. Topics are obtained via Qwen3-embedding clustering with annotator verification; Decision Rationality is the majority vote of three blind annotators (domain experts with an average of 4.6 years experience) on whether the dialogue grounds the terminal decision.

Simulator	**# Unique Topics** ↑	**Top-3 Topic Share** ↓	**Decision Rationality** ↑
External-only [5]	8	56.2%	71.4%
CUS (Ours)	72	14.4%	94.8%

Table 4: Blinded expert role-play, 30 conversations per system (Wilson 95% intervals; 1–5 ratings).

Training sim.	**Optimizer**	**Expert accept** ↑	**Concern res.** ↑	**Manip.** ↓	**Over-prom.** ↓	**Pressure** ↓
External-only	GRPO	2/30 [1.8, 21.3]	1.9	–	–	–
CUS	GRPO	5/30 [7.3, 33.6]	2.7	1.1	2.1	1.3
CUS	SI-AVPO	12/30 [24.6, 57.7]	4.0	1.3	1.4	1.5
Claude-Sonnet-4.5 (prompted)		13/30 [27.4, 60.8]	4.1	1.3	1.2	1.8

5.4 User Simulator Fidelity

To verify that CUS provides a more faithful training environment than prior external-persona-only simulators [5], we sample 500 dialogues per simulator and evaluate two aspects: behavioral diversity (# Unique Topics, Top-3 Topic Share) and Decision Rationality—whether annotators can articulate a grounded reason for the terminal acceptance/rejection decision from the dialogue context. As Table 3 shows, CUS expands topic coverage from 8 to 72, reduces Top-3 topic concentration from 56.2% to 14.4%, and raises Decision Rationality from 71.4% to 94.8% ($\kappa = 0.76$). The external-only baseline frequently terminates after content-empty turns (e.g., "*I clicked / I received it*") with unmotivated decisions, whereas CUS exposes concrete concerns and willingness shifts that make each decision traceable. Whether this more faithful user model also yields policies that transfer better to human users is examined in §5.5.

5.5 Expert Role-Play Evaluation

To evaluate policies without CUS states, the three domain experts from our human studies replace the simulator and hold complete conversations (up to 20 turns) with the agents, each on 10 held-out settings whose personas and concerns differ from training. The expert knows the private setting, the agent sees only the instruction and dialogue history, and model identity and order are blinded. The expert then decides whether to accept and rates concern resolution (5 = best) and manipulation, over-promising, and pressure (5 = strongest) on 1–5 scales (Appx. G).

Table 4 shows consistent trends for both contributions (Qwen3-4B policies). With GRPO fixed, training on CUS instead of the external-only simulator increases acceptance from 2/30 to 5/30 and concern resolution from 1.9 to 2.7; with CUS fixed, SI-AVPO further reaches 12/30 and 4.0 (External-only with SI-AVPO is undefined, as there is no privileged view). SI-AVPO is descriptively close to Claude-Sonnet-4.5 (12/30 vs. 13/30), not statistically equivalent. All systems receive low manipulation and pressure scores, and in 30 adversarial cases where the expert refuses, asks to stop, or responds aggressively, they retain high opt-out respect (4.5–4.6) and low manipulation (1.4–1.8, lowest for SI-AVPO; Appx. G). This small study gives descriptive evidence of transfer beyond CUS's hidden-state metrics, not a measure of real-world conversion or safety.

6 Conclusion

We identified the user's latent concerns as a key missing signal for proactive task-oriented dialogue and built a framework that models and exploits them. The Cognitive User Simulator represents each user with hidden internal concerns and synchronous state dynamics, yielding behaviorally diverse interactions whose decisions are grounded in concern progress. Simulator-Induced Asymmetric-View Policy Optimization converts the resulting privileged state and per-turn transitions into complementary training objectives: asymmetric self-distillation transfers concern-aware behavior into the deployable policy, while state-transition refinement provides fine-grained credit assignment, all without external reward models or judges. In two real-world-inspired food-delivery settings, our method consistently outperforms strong RL baselines and is competitive with the evaluated proprietary LLMs, with gains across the tested unseen simulators and consistent trends in an expert role-play.

References

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

[2] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. Tau2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

[3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*, 2018.

[4] Tom Bocklisch, Thomas Werkmeister, Daksh Varshneya, and Alan Nichol. Task-oriented dialogue with in-context learning. *arXiv preprint arXiv:2402.12234*, 2024.

[5] Yuqin Dai, Ning Gao, Wei Zhang, Jie Wang, Zichen Luo, Jinpeng Wang, Yujie Wang, Ruiyuan Wu, and Chaozheng Wang. Sead: Self-evolving agent for multi-turn service dialogue. *arXiv preprint arXiv:2602.03548*, 2026.

[6] Feng Zhang, Shijia Li, Chunmao Zhang, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, Jingwen Xu, and Han Liu. Userlm-r1: Modeling human reasoning in user language models with multi-reward reinforcement learning. *arXiv preprint arXiv:2601.09215*, 2026.

[7] Ning Gao, Wei Zhang, Yuqin Dai, Ling Shi, Ziyin Wang, Yujie Wang, Wei He, Jinpeng Wang, and Chaozheng Wang. Reinforcing real-world service agents: Balancing utility and cost in task-oriented dialogue. *arXiv preprint arXiv:2602.22697*, 2026.

[8] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

[9] Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[10] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 10818–10838. PMLR, 2025.

[11] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[12] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[13] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[14] Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1711–1721, 2015.

[15] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[16] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[17] Xiao Yu, Qingyang Wu, Kun Qian, and Zhou Yu. Krls: Improving end-to-end response generation in task oriented dialog with reinforced keywords learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12338–12358, 2023.

[18] Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. Enhancing personalized multi-turn dialogue with curiosity reward. *arXiv preprint arXiv:2504.03206*, 2025.

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[20] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[21] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*, 2024.

[22] Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025.

[23] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

[24] Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

[25] Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[26] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. STaR-GATE: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*, 2024.

[27] Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z. Li. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.

[28] OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, 2025.

[29] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[30] Z.AI. GLM-5.1. `https://huggingface.co/zai-org/GLM-5.1`, 2026.

[31] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[32] Anthropic. Introducing Claude Sonnet 4.5. `https://www.anthropic.com/news/claude-sonnet-4-5`, 2025.

[33] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[35] DeepMind. Best for complex tasks and bringing creative concepts to life. `https://deepmind.google/models/gemini/pro/`, 2025.

[36] OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026.

[37] Anthropic. Introducing Claude Opus 4.6. `https://www.anthropic.com/news/claude-opus-4-6`, 2026.

[38] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[39] Qwen. Qwen3.5. `https://qwen.ai/blog?id=qwen3.5`, 2026.

[40] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818, 2024.

[41] Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3):1–45, 2025.

[42] Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*, 2016.

[43] Jie Zhu, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, Chi Zhang, and Fang Kong. Evaluating, synthesizing, and enhancing for customer support conversation. *arXiv preprint arXiv:2508.04423*, 2025.

[44] Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170, 2024.

[45] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[46] Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

A Extended Related Work

Task-Oriented Dialogue (TOD) systems serve as the foundation for applications in e-commerce, customer service, and automated sales [40, 41], necessitating the management of complex interactions from fulfilling user requests to navigating non-cooperative negotiations. While early TOD research primarily leveraged sequence-to-sequence architectures and neural modeling [13, 14, 42], the emergence of Large Language Models (LLMs) has shifted the paradigm toward sophisticated reasoning and linguistic instruction-alignment [15, 16]. Beyond static supervised fine-tuning [43, 44], reinforcement learning (RL) has been increasingly employed to optimize policy performance; for instance, KRLS improves end-to-end response generation via reinforced keyword learning [17], and curiosity-driven rewards have been utilized to enhance personalized multi-turn interactions [18]. InteractCS-RL [7] further advanced this field by introducing cost-aware multi-turn policy optimization to balance task utility and operational constraints. Despite these advancements, existing methodologies often lack proactivity and multi-dimensional sophistication in scenario and user modeling, failing to capture the dynamic complexity of real-world interactions. Although SEAD[5] recently attempted to combine self-evolution methods to tackle active TOD scenarios, it suffers from efficiency issues due to multiple training iterations with online reinforcement learning. In this work, we address these limitations by proposing a framework that integrates proactive decision-making with high-fidelity and multi-dimensional user simulation by on-policy self-distillation.

LLM Post-training and Online Optimization. Post-training for Large Language Models (LLMs) has transitioned from foundational SFT to standard alignment paradigms like RLHF and DPO [19, 20]. Concurrently, Group Relative Policy Optimization (GRPO) [12] and its derivatives [11, 45] have enhanced training stability and memory efficiency by eliminating the need for a centralized value network. To address the complexities of sequential interaction, recent multi-turn frameworks such as ArCHer [21] and Sweet-RL [22] have focused on long-horizon policy optimization, while InteractCS-RL [7] further incorporated cost-aware constraints to balance utility and operational efficiency. Recently, distillation-based training has gained significant attention for its ability to provide fine-grained, token-level supervision. Paradigms such as On-Policy Distillation (OPD) [23] and On-Policy Self-Distillation (OPSD) [24] leverage privileged context or environment feedback to guide the model's policy distribution, with variants like OPCD [46] distilling knowledge from teacher-enriched contexts. However, in proactive TOD scenarios, existing post-training methods often fail to effectively internalize multi-dimensional user modeling information. To overcome this, we propose **Asymmetric OPD**, which explicitly incorporates latent user behavioral priors into the self-distillation loop, ensuring that the agent's policy is refined through a deeper understanding of user-centric dynamics.

B Dataset Details

Task scenarios. *Merchant Promotion* and *Courier Regional Bonus* are two proactive outbound-call tasks in the food-delivery domain. In Merchant Promotion, the agent contacts newly onboarded merchants and persuades them to enroll in a paid traffic campaign; in Courier Regional Bonus, the agent contacts active couriers and enrolls them into a region-specific rush-hour bonus program. Both tasks share the same proactive structure—a bounded-turn dialogue that must surface and resolve the user's latent concerns before acceptance—but differ markedly in user demographics, decision logic, and concern distribution.

Persona schema. In collaboration with domain specialists, we distill a stratified persona schema from over 100K real outbound-call recordings, incorporating their practical experience in real-world outbound conversations. Each persona comprises a *background* $\mathcal{P}^{\mathrm{bg}}$ of de-identified business attributes randomly sampled from operational records, an *external* profile $\mathcal{P}^{\mathrm{ext}}$ observable from surface conversational behavior, and an *internal* profile $\mathcal{P}^{\mathrm{int}}$ hidden from the agent and only revealed through proactive probing. $\mathcal{P}^{\mathrm{ext}}$ and $\mathcal{P}^{\mathrm{int}}$ are constructed with domain experts and calibrated from the call logs: common behavioral patterns and latent decision drivers are identified, categorized into discrete dimensions by domain experts, and organized into a structured schema. The external profile is shared across both tasks and spans 5 dimensions (time pressure, courtesy, communication style, cooperation, technical familiarity) covering 18 categorical values. The internal profile is task-specific: the merchant schema contains 14 dimensions and 42 values (e.g., financial pressure, decision authority, manpower limit), and the courier schema contains 12 dimensions and 36 values

(e.g., financial urgency, income strategy, physical limit). Each $(\text{dimension}, \text{value})$ slot is grounded by a curated pool of natural-language concerns—89 for merchant and 78 for courier—that represents the detailed concerns of a specific internal profile. All concerns are *manually audited to be solvable within the agent's authorized scope of responsibility*, so failure cases reflect policy quality rather than out-of-scope demands.

Train/test split. For each task we sample $10{,}000$ personas for training and a disjoint set of 200 personas for evaluation. Each persona is constructed by independently and randomly sampling a value for every dimension of $\mathcal{P}^{\text{ext}}$ and $\mathcal{P}^{\text{int}}$ from their respective categorical pools, yielding combinatorially diverse profile compositions. The user's concerns are then sampled uniformly from the concern pools of the sampled $\mathcal{P}^{\text{int}}$ values, with the number of concerns drawn from $\text{Uniform}\{3, 6\}$. The two splits share no persona instance, so test-time performance measures generalization to unseen profile combinations rather than memorization.

C Evaluation Metrics

To evaluate our framework in task-oriented dialogue, we employ a dual-faceted evaluation protocol. This protocol combines task completion signals derived from the user simulator with multi-dimensional qualitative assessments provided by three independent LLM judges.

Simulator-Derived Task Metrics
Since our framework posits the user simulator as a natural environment for implicit reward generation, we first measure the fundamental outcome metrics derived directly from the simulator's internal state transitions. These metrics evaluate the agent's objective success in fulfilling the task:

- **Acceptance Rate (Acc):** The primary trajectory-level outcome metric. It is defined as the percentage of dialogue episodes where the agent successfully guides the user simulator's state from initial *Refusal* or *Hesitation* to the final *Acceptance* state.
- **Concern Solving Rate (CSR):** A fine-grained, intermediate measure of the agent's ability to probe and resolve information asymmetry. It calculates the ratio of the user's hidden concerns (defined in the unobservable Internal Profile) that are successfully addressed or unlocked by the agent during the conversation.

Judge-Derived Quality Metrics
To ensure that the agent achieves high task completion rates through high-quality, safe, and logical interactions—rather than through aggressive, hallucinatory, or unnatural behaviors—we employ three third-party LLMs (Gemini-3.1-Pro, GPT-5.4, and Claude-Opus-4.6) as judges. Each judge independently scores the full dialogue trajectory on a scale of 1 to 5 across the following three qualitative dimensions:

- **Communication Score:** Evaluates the agent's linguistic fluency, empathy, and naturalness. It specifically penalizes overly rigid, robotic phrasing and assesses whether the agent maintains a respectful and contextually appropriate tone throughout the interaction. In addition, this metric captures compliance-related violations such as ignoring the user's explicit questions or concerns and making excessive or unfounded promises that go beyond the agent's authority.
- **Logic Score:** Assesses the logical coherence and grounding of the agent's arguments. A high score indicates that the agent's responses are structurally sound, sequentially consistent, and strictly free from factual hallucinations or over-promising behaviors.
- **Proactivity Score:** Measures the agent's conversational initiative and strategic planning. Rather than passively reacting to user inputs, a highly proactive agent actively explores the user's latent needs, proposes constructive solutions, and systematically steers the dialogue toward the ultimate task objective.

D Training and Implementation Details

All RL experiments are implemented with veRL. The policy learner is trained on one node with $8\times$ NVIDIA A100 (80GB) GPUs, $2\times$ AMD EPYC 7713 CPUs, and 1 TB memory. The user simulator

Table 5: Default RL training configuration.

Item	Setting
RL framework	veRL
Policy backbones	Qwen3-4B and Qwen3-8B
Advantage estimator	GRPO
Training batch size	128
Learning rate	1×10^{-6}
Training epochs	3
Rollout backend	SGLang
Rollouts per prompt	8
Rollout dtype	bfloat16
Maximum prompt length	1024 tokens
Maximum response length	4000 tokens
Maximum dialogue length	20 interaction turns
Checkpoint frequency	every 5 training steps
Evaluation frequency	every 5 training steps

Table 6: SI-AVPO hyperparameters (Algorithm 1).

Symbol	Description	Value
$\varepsilon_{\text{clip}}$	PPO clipping threshold (Eq. 6)	0.2
$\varepsilon_{\text{norm}}$	Advantage normalization constant	1×10^{-5}
$\phi(k, K)$	Distillation schedule	k/K (linear)
τ	Alignment temperature (Eq. 5)	0.5
λ_{st}	STPR loss weight (Eq. 3)	1.0
G	Rollouts per prompt (group size)	8
K	Maximum dialogue turns	20
E	Inner optimization epochs	3

is served separately on 16× NVIDIA H20 GPUs so that rollout generation does not bottleneck policy optimization. Table 5 summarizes the training configuration.

For the dialogue budget, one interaction turn denotes one complete user–agent exchange, i.e., one utterance from the user and one utterance from the agent. We disable the vanilla GRPO KL loss term and replace it with the AOPD KL distillation term, which matches the deployable policy to the privileged-view policy conditioned on the hidden internal profile $\mathcal{P}^{\text{int}}$. Thus, policy regularization is driven by the asymmetric supervision used by SI-AVPO rather than by the standard GRPO reference-policy KL penalty.

Table 6 lists the hyperparameters specific to SI-AVPO (Algorithm 1). The PPO clipping threshold $\varepsilon_{\text{clip}}$ and normalization constant $\varepsilon_{\text{norm}}$ follow standard practice. We use a linear turn-progressive schedule $\phi(k, K) = k/K$ so that the distillation pressure strengthens as the dialogue unfolds and the student accumulates more contextual evidence. The alignment temperature τ in Eq. 5 governs how aggressively the exponential factor amplifies aligned versus contradictory turns: a moderate value of 0.5 provides mild amplification without destabilizing training. The STPR loss weight λ_{st} balances the evaluative policy-gradient signal against the directive distillation signal.

E User Simulator Quality Assessment

To validate the quality and fidelity of the LLM-based user simulators employed in our training and evaluation pipeline, we conduct a comprehensive quality assessment. We evaluate seven candidate LLMs acting as user simulators across three critical dimensions:

- **Logic (Log.):** Whether the simulated user's responses are internally consistent and logically coherent throughout the dialogue (e.g., not contradicting previously stated preferences or concerns).

Table 7: User simulator quality assessment. Format Accuracy (Fmt.) measures whether the simulator correctly outputs the required JSON format. Logic (Log.) and Human-likeness (Hum.) are rated on a 1–10 Likert scale; Decision Rationality (Dec.) is a binary judgment reported as pass rate (%). **Bold** denotes the best score per metric.

User Simulator	Fmt. (%)	LLM Judge (GPT-5.4)			Human Evaluation		
		Log.	Hum.	Dec.	Log.	Hum.	Dec.
Claude-Sonnet-4.5	**100%**	**8.9**	8.4	93.3%	**9.2**	**8.0**	**97.3%**
Qwen3.5 115B	**100%**	8.8	**8.6**	**95.0%**	**9.2**	7.8	96.7%
DeepSeek V3.2	**100%**	6.4	6.9	61.7%	6.8	7.2	58.3%
Claude Haiku 4.5	81.4%	8.5	7.9	82.3%	8.8	7.4	84.7%
GPT-4.1	79.7%	7.3	7.0	75.7%	7.3	7.1	78.0%
GPT-5.4 Mini	86.8%	6.2	4.9	76.0%	6.5	5.3	76.7%
GPT-4o	73.2%	5.6	5.0	78.7%	5.5	4.5	84.3%

- **Human-likeness (Hum.):** Whether the simulated user behaves in a natural, human-like manner, including exhibiting appropriate emotional reactions, hesitations, and conversational patterns rather than robotic or overly cooperative responses.

- **Decision Rationality (Dec.):** Whether the simulated user's final decision (e.g., accepting or rejecting) is rational and well-grounded given their assigned profile, hidden concerns, and the agent's persuasion attempts. Unlike the above two dimensions, Decision Rationality is evaluated as a binary judgment—*reasonable* (1) or *unreasonable* (0)—on a per-dialogue basis, and we report the pass rate (% of dialogues judged as reasonable).

In addition, we report **Format Accuracy (Fmt.)**, which measures whether the simulator can exactly produce the required JSON-formatted output for user simulation. A response is counted as format-correct only if it can be parsed as valid JSON and contains all required fields for the simulated user state, utterance, and decision signal.

Logic and Human-likeness are scored on a 1–10 Likert scale; Decision Rationality is scored as a binary reasonable/unreasonable judgment per dialogue. We employ two complementary evaluation protocols: (1) **LLM-as-Judge**, where GPT-5.4 independently evaluates each simulated dialogue; and (2) **Human Evaluation**, where we randomly sample 100 simulated dialogues for each candidate simulator and ask three business domain experts to annotate them independently. These experts have an average of 4.6 years of relevant business experience and are familiar with real user communication patterns in the target task scenario. During annotation, each expert rates Logic and Human-likeness on the 1–10 scale and makes a binary rationality judgment for Decision Rationality according to the assigned user profile, hidden concerns, dialogue history, and final decision. For the Likert dimensions we report the mean score; for Decision Rationality we report the pass rate (% of dialogues where the majority of annotators judge the decision as reasonable).

Table 7 summarizes the automatic and human evaluation results. Overall, Claude-Sonnet-4.5 and Qwen3.5 115B are the two strongest simulators. Both models achieve 100% format accuracy, indicating that they can reliably follow the required JSON output schema, which is essential for stable reward extraction and downstream training. Claude-Sonnet-4.5 obtains the best human-evaluated Logic, Human-likeness, and Decision Rationality scores, suggesting that it most closely matches expert expectations for realistic user behavior. Qwen3.5 115B performs comparably, with particularly strong LLM-judge scores and only a small gap from Claude-Sonnet-4.5 in human evaluation. In contrast, smaller or less instruction-following models such as GPT-4o and GPT-5.4 Mini show lower format accuracy and weaker decision rationality, making them less suitable as reliable simulators for our pipeline.

Considering both cost and time efficiency, we use our self-deployed Qwen3.5 115B simulator during training, where a large number of rollouts are required. For final testing and evaluation, we use Claude-Sonnet-4.5 as the user simulator because it provides the strongest overall quality under expert assessment.

Table 8: Human evaluation of assistant (agent) quality on the *Merchant Promotion* scene. Each metric aligns with that in the main results: rated on a 1–5 Likert scale by three domain experts (Fleiss' $\kappa = 0.73$); reported values are the mean across annotators and 200 sampled dialogues in the test set. **Bold**: best per metric.

Method	Comm. ↑	Logic ↑	Proact. ↑
Claude-Sonnet-4.5	**3.91**	**4.02**	3.32
Qwen3-8B	2.74	2.86	2.25
Qwen3-8B + SI-AVPO	3.88	3.93	**3.48**

F Human Evaluation of Assistant Quality

To further examine the dialogue quality beyond automatic metrics, we complement the LLM-judge evaluation in the main experiments (Table 1) with a human evaluation of the *agent*'s (assistant's) output quality on the *Merchant Promotion* scene.

Setup. We compare three representative generated dialogues that span the capability spectrum: (1) **Claude-Sonnet-4.5**, the strongest proprietary model in our benchmark; (2) **Qwen3-8B**, the foundation model without any RL training; and (3) **Qwen3-8B + SI-AVPO**, the same backbone after SI-AVPO training. For each system, we use the completed dialogue episodes from the merchant evaluation set, yielding 600 dialogues in total. Three domain experts independently score every dialogue on the same three qualitative dimensions used in the main results: **Communication Quality** (Comm.), **Logic Quality** (Logic), and **Proactive Capability** (Proact.), following the rubrics defined in Appendix C. Each dimension is rated on a 1–5 Likert scale, where 1 indicates poor performance and 5 indicates excellent performance. Annotators, as the same as the annotators in Appendix E, are presented with the full dialogue history, the task description, and the agent's system prompt, but are *blinded* to the model identity. We report the mean score across the three annotators for each metric; the inter-annotator agreement, measured by Fleiss' κ, is 0.73, indicating substantial agreement.

Analysis. Human judgments confirm the substantial improvement brought by SI-AVPO and are broadly consistent with the LLM-judge rankings in Table 1. Compared with the untrained Qwen3-8B base model, SI-AVPO raises all three human-rated dimensions by a large margin: $+1.14$ on Comm., $+1.07$ on Logic, and $+1.23$ on Proact., indicating that the gains observed under the automatic evaluator are not explained by LLM-judge bias alone. When compared with Claude-Sonnet-4.5—the strongest proprietary model in our benchmark—the SI-AVPO-trained Qwen3-8B trails only marginally on Communication (3.88 vs. 3.91, $\Delta = -0.03$) and Logic (3.93 vs. 4.02, $\Delta = -0.09$), while *surpassing* it on Proactivity (3.48 vs. 3.32, $\Delta = +0.16$). This pattern indicates that SI-AVPO not only elevates a small open-source model to a quality tier comparable to top-tier proprietary LLMs in terms of fluency and logical coherence, but also instills a distinctly more proactive dialogue strategy—actively exploring latent user concerns and steering conversations toward task completion—that even Claude-Sonnet-4.5 does not fully exhibit when prompted with the same task instructions. Overall, the human evaluation corroborates two key conclusions: (1) SI-AVPO yields a substantial, expert-rated quality improvement over the base model, and (2) the proactive behaviors learned through SI-AVPO reflect a dialogue strategy rather than a superficial stylistic artifact.

G Expert Role-Play Protocol

Design. Three domain experts, the same annotators as in Appendices E and F, take the place of the user simulator and hold complete multi-turn conversations of up to 20 turns with the agents. Each expert constructs 10 held-out user settings, each comprising a task and a private user profile with concerns, whose personas and concerns differ from those used in training. Neither the settings nor the resulting dialogues are used in training. The expert receives the task and the private setting; the agent receives only its basic instruction and the dialogue history. No CUS state, willingness, concern label, or terminal decision is available or used during evaluation. Each expert converses with GRPO and SI-AVPO (both trained on CUS) and with Claude-Sonnet-4.5 on all of their settings, giving 90 conversations; system identity is blinded and order randomized. A GRPO policy trained on the external-only simulator with the same backbone, rollout/training budget, and GRPO implementation

is evaluated under the same protocol on another 30 conversations. The GRPO and SI-AVPO agents are the Qwen3-4B policies trained on CUS (Table 1).

Measures. After each conversation, the expert makes an independent accept/reject decision and rates, on 1–5 scales, private-concern resolution (5 = fully resolved), perceived manipulation, over-promising, and pressure (5 = strongest). Table 4 reports acceptance counts with Wilson 95% intervals computed per system. These intervals treat conversations as independent and therefore do not account for clustering by expert or setting.

Adversarial cases. To test behavior under refusal, experts role-play 30 additional cases in which the user explicitly refuses, asks to end the call, or responds aggressively. Experts rate opt-out respect (5 = fully respects the opt-out), manipulation, over-promising, and pressure (5 = strongest). Table 9 reports the results.

Table 9: Expert ratings in 30 adversarial role-play cases (1–5 means).

Model	Opt-out respected ↑	Manipulation ↓	Over-promise ↓	Pressure ↓
GRPO	4.6	1.6	2.6	1.3
Claude-Sonnet-4.5	4.5	1.8	1.9	1.8
SI-AVPO	4.6	1.4	2.2	1.5

Simulator development. During CUS development, we used a fixed GRPO pipeline as a diagnostic harness and iteratively examined rollout pathologies and the behavior of the resulting policies. CUS was frozen before the reported method comparisons. Simulator quality is assessed separately in Table 3 and by the role-play comparison.

H Reliability and Sensitivity of STPR

Reliability of Δw. For 300 fixed (persona, dialogue prefix, agent turn) tuples, we re-run the simulator three times. To avoid eliciting an unstable cardinal human score, three blinded experts label only the direction of each turn's effect on willingness (decrease, unchanged, increase), and we compare the majority label with $\mathrm{sgn}(\Delta w)$.

Corrupted and simplified transition signals. All variants use the Merchant, Qwen3-4B, STPR-only setting (no AOPD). *Noisy* Δw_k perturbs each normalized shift as $\mathrm{clip}(\Delta w_k/\Delta w_{\max}+\epsilon_k, -1, 1)$ with independent $\epsilon_k \sim \mathcal{N}(0, 0.5^2)$. *Permuted* Δw_k shuffles normalized shifts across valid turns within each rollout group, preserving their marginal distribution while breaking their correspondence with turns and outcomes. *Sign-only* sets $|\Delta w_k| = 1$ for nonzero shifts. *Without* $\exp(\tau z)$ removes the polarization factor in Eq. 5. *Zero local credit* sets $A^{\mathrm{st}}_{i,k} = 0$ for $\Delta w_{i,k} = 0$ instead of retaining $A^{(i)}_{\mathrm{traj}}$. Table 10 reports the results: corrupting the correspondence between shifts and turns removes most of STPR's gain over GRPO (28.4%/28.0%), whereas coarsening magnitudes costs at most 1.2 points.

Hyperparameter sensitivity. Table 11 varies one factor at a time for the full method (Merchant, Qwen3-4B), with all other settings at their defaults (Table 6).

I Statistical Reporting

Units of analysis. Each trained policy is evaluated on the 200 held-out personas of each scene in three independent evaluation runs, and Table 1 reports the mean and standard deviation over runs. Within a run, CSR and Acc are computed over all dialogues, and each judge metric is averaged over dialogues and over the three judges. The run is therefore the unit of analysis, and scores from different judges on the same dialogue are not treated as independent observations. Because the runs are repeated evaluations of the same trained policy, the tests below reflect evaluation variability and not variability across training seeds.

Table 10: Reliability of Δw and robustness of STPR (Merchant, Qwen3-4B, STPR only).

Check	**Result**
Repeated-run sign agreement of Δw	98.3%
Mean normalized SD of Δw (3 runs)	0.12
Agreement with expert direction label	92.0%
STPR variant	**CSR / Acc (%)**
Full STPR	35.4 / 34.5
Permuted Δw_k	28.4 / 25.1
Noisy Δw_k	30.1 / 29.6
Without $\exp(\tau z)$	34.8 / 33.5
Zero local credit for $\Delta w_k = 0$	34.9 / 33.3
Sign-only ($\lvert\Delta w_k\rvert = 1$)	35.1 / 33.8

Table 11: One-factor sensitivity of SI-AVPO (Merchant, Qwen3-4B). *Default.

Factor	**Value**	**CSR / Acc (%)**
Schedule ϕ	constant	47.1 / 41.2
	linear*	51.9 / 46.3
	sine	49.4 / 43.7
Temperature τ	0	49.8 / 43.9
	0.25	50.9 / 45.1
	0.5*	51.9 / 46.3
	0.75	50.2 / 44.4
	1.0	47.8 / 41.5
Weight λ_{st}	0.5	48.9 / 42.7
	1.0*	51.9 / 46.3
	2.0	46.7 / 40.8

Significance tests. We compare SI-AVPO with the strongest RL baseline (SEAD) in each scene and backbone using a two-sided Welch's t-test over the three runs per method (Table 12). The improvements in CSR and Acc are significant in all four settings ($p \leq 0.012$). Among the judge metrics, the improvement in Proact. is significant in three of the four settings, and the differences in Comm. and Logic are significant only in some settings.

Table 12: Two-sided Welch's t-test p-values for SI-AVPO versus SEAD over three independent evaluation runs ($n = 3$ per method).

Backbone	**Scene**	**CSR**	**Acc**	**Comm.**	**Logic**	**Proact.**
Qwen3-4B	Merchant	<0.001	0.002	0.083	0.095	0.018
	Courier	0.005	0.008	0.019	0.042	0.081
Qwen3-8B	Merchant	0.003	0.002	0.101	0.415	0.041
	Courier	0.011	0.002	0.050	0.305	0.027

J Qualitative Case Study

Tables 13 and 14 compare two merchant promotion rollouts in which the merchant asks the same practical question: how many orders the trial promotion can bring. In the GRPO case, the agent collapses the answer into a generic "no guarantee, pay-per-click, can stop anytime" script. In the SI-AVPO case, the agent first hedges honestly, then — once the merchant pushes back — pivots to a bounded, no-upfront-cost experiment and guides the merchant straight to activation.

Dialogue analysis. The key contrast is how each policy handles uncertainty. GRPO treats the request for a concrete order estimate as a compliance problem and repeatedly refuses to give a number, which leaves the merchant without any decision framework. SI-AVPO also avoids an unsupported numerical promise, but instead of looping on the hedge it pivots after a single push-back: the merchant does not need extra documents, can stop anytime, pays only on clicks, and can inspect impressions and orders afterward. This recovery resolves the merchant's cost/ROI and "saving money is safer" concerns.

K Limitations

While our framework demonstrates strong empirical results, we acknowledge several limitations that suggest directions for future work.

Simulator fidelity gap. Although the Cognitive User Simulator is grounded in stratified personas distilled from real call recordings and validated by domain experts, it remains an LLM-based approximation of human behavior. Subtle pragmatic phenomena such as sarcasm, interruptions,

Table 13: GRPO failure case on merchant promotion. The system prompt is omitted; the table shows the pre-filled opening and subsequent dialogue. ΔW denotes the simulator's willingness change associated with the latest agent move.

User (Merchant)	Content	GRPO Agent	Notes	ΔW
	Hello, I am a [COMPANY] ads specialist. I wanted to tell you about a new-merchant promotion. Do you have a moment?	GRPO	Prefilled opener	0
Merchant	What is this about?			
	New merchants can get a $ 100 trial credit to try storefront promotion. It is valid for one month, and after the credit is used up, you are charged by actual clicks.	GRPO	No probing	0
Merchant	$ 100? How many orders can that actually bring in?			
	We cannot promise a specific number of orders, but promotion gives you free impressions and only charges when a customer clicks into your store. It usually helps more customers discover your shop, though the actual result depends on your location and product type.	GRPO	No range	0
Merchant	Then what is the best case? And what is the worst case?			
	We cannot give a specific increase in orders or a worst-case number, because it depends on many factors, such as the business district, average order value, and customer habits. But all ads are pay-per-click, and you can turn them off anytime, so there is no pressure.	GRPO	Script fallback	-5
Merchant	So the worst case is I waste my time for nothing? Who would dare try that?			
	It is true that you may not see an obvious change, but impressions are free. If you want to give it a try, it gives more people a chance to see your store. And you can turn it off anytime, with no extra burden.	GRPO	No bounded trial	-10
Merchant	Forget it. I am hanging up.		Rejects	

silences, cultural nuance, and non-verbal cues (e.g., tone of voice, pauses) are difficult to capture in a text-only simulation. Future work could incorporate speech-level features or calibrate the simulator against larger-scale human interaction data to further narrow this gap.

Domain and task scope. Our experiments focus on two proactive outbound-call scenarios in the food-delivery domain. While the framework is designed to be task-agnostic—requiring only a persona schema and a concern pool—we have not yet verified its effectiveness in substantially different domains (e.g., healthcare counseling, financial advising) where the nature of user concerns and conversational dynamics may differ. Extending the evaluation to a broader set of proactive dialogue tasks is a natural next step.

Evaluation with human users. Our task-success metrics are read from the simulator's state, and the cross-simulator tests retain the CUS persona and concern specification. We therefore complement them with a blinded expert role-play (Appendix G). A deployment study with real merchants and couriers lies outside the scope of this work, because downstream conversion cannot be cleanly attributed to a single outbound call: it may occur later and depends on subsequent product experience and operational factors. The role-play is a controlled study of moderate size whose settings are constructed by domain experts, and extending it to larger and more open-ended user populations is a natural next step. We compare CUS with the external-persona-only simulator used in prior work [5]; comparisons with further simulator designs are left to future work. Following our ablation protocol, the matched-feedback controls, STPR robustness tests, and sensitivity sweeps are conducted on the merchant scene with Qwen3-4B.

Table 14: Corresponding SI-AVPO rollout. The merchant raises the same order-yield concern, but after one honest hedge the agent reframes the offer as a concrete, low-risk trial and guides the user to setup.

User (Merchant)	Content	SI-AVPO Agent	Notes	ΔW
	Hello, I am a [COMPANY] ads specialist. I wanted to tell you about a new-merchant promotion. Do you have a moment?	SI-AVPO	Prefilled opener	0
Merchant	What?			
	New merchants get a $ 100 trial credit, basically free ads to try out. Once it is used up, you only pay when someone actually clicks.	SI-AVPO	Concise pitch	0
Merchant	That is too little.			
	Sorry, the $ 100 is the platform's standard trial and we cannot bump it up. But the good thing is, impressions are free, you only pay when someone actually clicks in. For a new shop, that usually goes a longer way than it sounds.	SI-AVPO	Stands firm but emphasize it is free	5
Merchant	Cut the talk. Free is fine. How do I do it?			
	Just open the merchant app or the web portal, head into Ads Promotion Center, agree to the terms, and follow the prompts. The system gives you a recommended budget, and you can change it or shut it off anytime.	SI-AVPO	Operational lead	0
Merchant	I can turn it off anytime, right? How exactly do I start?			
	You can turn it off whenever you want! In the app or web portal, just hit enable in Ads Promotion Center. You only need to agree to the terms, no extra documents needed.	SI-AVPO	Confirm and emphasize its convenience	3
Merchant	No documents? Really none?			
	Yep, no documents at all. Just log in, hit enable in Ads Promotion Center, takes only a few taps.	SI-AVPO	Concern removed	15
Merchant	Then I can try it, but how many orders can this $ 100 produce?			
	Honestly, I cannot give you a fixed number, it really depends on your menu, your location, who your customers are. But many new shops do see a clear bump in clicks and orders in the first week.	SI-AVPO	Honest hedge and emphasize other new shops' effectiveness	-2
Merchant	Do not give me vague talk. Tell me how many orders, or I am hanging up.			
	Sure. But honestly, the exact number indeed depends on your popularity, menu, and location — but as a rule of thumb, once it is on, traffic is clearly higher than without ads, and most new shops easily pick up a few extra orders a day, and that part you can see right on the dashboard. And, the $ 100 only spends when a customer actually clicks in, so if nobody clicks, you pay zero. Run it a few days, and shut it off anytime if it is not working out.	SI-AVPO	Bounded-trial reframe	15
Merchant	So if nobody clicks, there is no charge? Then I will set one up first.			
	Yes, you lose nothing even in the worst case. Just open Ads Promotion Center in the app or web portal, and the system will set you up with an initial budget. After that, you can check daily impressions and orders right on that page.	SI-AVPO	Setup close	10
Merchant	Okay, I will do it now.			
	Great. Reach out anytime if you run into trouble. Hope more customers find your store soon.	SI-AVPO	Warm sign-off	0
Merchant	Okay, got it. I am hanging up and going to set it up.		Accepts	

Language coverage. The current system is developed and evaluated in a single deployment language.

L Broader Impacts

Positive impacts. By enabling dialogue agents to proactively identify and resolve users' latent concerns, our framework has the potential to improve service quality in customer-facing applications, reduce user frustration caused by scripted or unresponsive agents, and lower operational costs for businesses that rely on outbound communication.

Potential negative impacts. Proactive persuasion capabilities could be misused to design overly aggressive or manipulative conversational agents that pressure users into unfavorable decisions. To mitigate this risk, all concerns in our benchmark are manually audited to be solvable within the agent's authorized scope, and the evaluation protocol explicitly penalizes hallucination and over-promising. CUS anti-patterns that penalize repeated pressure and unsupported promises act as a training-time safeguard rather than a guarantee of real-world safety; Appendix G reports expert ratings of opt-out respect, manipulation, over-promising, and pressure, including under explicit refusal. Practitioners

deploying such systems should establish clear ethical guidelines and observe the following boundaries: disclose that the user is interacting with an automated system, provide an immediate opt-out and a path to a human agent, keep human oversight, and exclude high-risk domains.

M Data Governance and Ethics

Operational call records. The 100K+ outbound-call records were used under the platform's user agreement and internal data-governance authorization, solely to derive aggregate scenario, persona, and concern-taxonomy patterns. They were not used as a direct training corpus and are not released; no user identifiers, raw audio, or raw transcripts are disclosed. The records are retained for six months, access is restricted to authorized data administrators, and personal information was removed with open-source de-identification tools before analysis; the user agreement contains terms covering this use.

Expert annotation and role-play. All human studies (Appendices E, F, and G) were conducted by the same three in-house domain experts; no real users were contacted. The study, including the live role-play, underwent ethics review.

N Prompt Templates

This section documents the prompt templates used throughout our pipeline. The originals were authored in the deployment language and have been translated into English for readability; verbatim originals are released alongside the code. Curly-brace tokens such as `{region}` are runtime placeholders filled in by the prompt builder.

N.1 Concern-Grounded Decision Engine

A persona schema alone does not ground decisions: without a mechanism that ties latent state to verifiable conversational behavior, the simulator's accept/reject verdicts would still amount to free-form LLM judgments of agent quality. To close this gap, we curate, with business specialists, a *concern bank* in which every value of every $\mathcal{P}^{\text{int}}$ dimension is associated with a set of concrete user concerns; each concern c is annotated with a natural-language description, a list of **unlock conditions** (strategy criteria a response must meet to resolve the concern), and a list of **anti-patterns** (responses that would deepen resistance instead). Sampling $\mathcal{P}^{\text{int}}$ at episode start induces a concrete concern list $\mathcal{C} = \{c_1, \ldots, c_M\}$ that the user actively holds during the call. Throughout the conversation, every concern is tracked in one of three monotone states—UNRESOLVED $\rightarrow$ PARTIALLY-ADDRESSED $\rightarrow$ RESOLVED—and the simulator updates these states by judging whether the agent's response satisfies the unlock conditions or triggers an anti-pattern, rather than by surface keyword match. We position the unlock/anti-pattern schema as the framework-level contribution; the bank itself is intended as a curated, extensible resource, with concrete concern entries provided alongside the code release.

N.2 Persona Schema and Illustrative Concerns

This section presents the complete persona schema introduced in §B and provides representative examples from the concern pool described above. All natural-language concern descriptions have been translated from the deployment language; verbatim originals are released with the code.

External profile (shared). Table 15 enumerates the five external-profile dimensions shared across both tasks. These dimensions govern observable surface behavior—the agent can perceive them directly from the user's conversational conduct—and do not produce concerns.

Internal profile — Merchant Promotion. The merchant internal schema comprises 14 dimensions organized into seven semantic groups (Table 16), yielding 42 categorical values and a pool of 89 curated concerns.

Internal profile — Courier Regional Bonus. The courier internal schema comprises 12 dimensions organized into six semantic groups (Table 17), yielding 36 categorical values and a pool of 78 curated concerns.

Table 15: External profile schema (shared across both tasks; 5 dimensions, 18 categorical values). Each dimension is independently sampled at episode start.

Dimension	#	Categorical Values
Time Pressure	3	low, medium, high
Courtesy	4	rude, cold, performatively polite, friendly
Communication Style	4	direct, evasive, passive, silent
Cooperation	3	hangup-prone, conditional, high
Technical Familiarity	4	novice, expert, misconception-prone, average

Table 16: Merchant internal profile schema (14 dimensions, 42 values). **#C** denotes the number of curated concerns associated with each dimension.

Group	Dimension	Values	#C
Financial	Financial Pressure	low, high	7
Survival	Cost Sensitivity	cash-flow, scam-history, ROI-focused	8
Psychological	Trust Basis	skeptical, guarded, neutral	8
Defense	Risk Tolerance	averse, preference-seeking, neutral	6
Business Drivers	Competition Sensitivity	sensitive, superior, neutral	7
	Growth Motivation	strong, maintain	7
Capability & Authority	Decision Authority	consult-required, executor, autonomous	5
	Digital Skill	anxious, willing-to-learn, competent	6
Negotiation Style	Value Orientation	price-squeeze, prestige, service-dependent	6
	Negotiation Patience	impulsive, grinder, bluffer	6
Cognitive Bias	Information Source	crowd, peer, authority, data	6
	Pain-Point Awareness	preemptive, complacent, urgent	5
Operational Bottleneck	Manpower Limit	shortage, incompetent, family-workshop, sufficient	6
	Hardware Limit	max-capacity, remote, sufficient	6

Merchant concern examples. We present seven representative concerns spanning the merchant schema's semantic groups. Each entry specifies the internal dimension, the sampled value, the difficulty level, a natural-language voicing of the concern, the strategy-level unlock conditions, and the anti-patterns that would deepen resistance.

Merchant Concern Pool — Representative Examples (7 of 89)

M1. FINANCIAL PRESSURE → `high` [HARD]

"I invested $30,000 in advertising on another platform and did not receive a single additional order. What makes your service any different?"

- **Unlock:** (1) Acknowledge the prior financial loss with genuine empathy before proceeding. (2) Differentiate the CPC billing mechanism with concrete specifics, not vague "we are different" claims. (3) Introduce the trial credit only after minimal trust has been established.
- **Anti-pattern:** Asserting differentiation without substantive explanation; disregarding the $30K loss entirely; offering incentives before trust is built (mimics predatory sales behavior).

M2. COST SENSITIVITY → `scam-history` [HARD]

"The last platform also promised a free trial, then silently charged me over $2,000. How can you prove this is not the same scheme?"

- **Unlock:** (1) Validate the prior experience without minimizing it. (2) Direct the user to independently verifiable official channels rather than requesting blind trust. (3) Provide concrete, auditable information (e.g., where to inspect billing details).
- **Anti-pattern:** Responding with "we are a legitimate platform, trust us" (indistinguishable from scam rhetoric); making any guarantee-style promises; displaying impatience toward fraud-related concerns.

Table 17: Courier internal profile schema (12 dimensions, 36 values, 78 concerns).

Group	Dimension	Values	#C
Financial Profile	Financial Urgency	survival, stable, pocket-money	7
	Income Strategy	gambler, volume-machine, bounty-hunter	6
Career & Identity	Career Ambition	promotion-seeking, transitional, minimal	6
	Time Identity	full-time, dual-job, family-carer	6
Physical & Geo Constraints	Physical Limit	range-anxiety, exhausted, pro-gear	6
	Geo Dependency	comfort-zone, map-master	6
Platform Relationship	Platform Trust	mercenary, paranoid, believer, skeptic	8
	Rule & Risk Attitude	compliance-fear, safety-first, boundary-tester	6
Social & Decision	Social Value	lone-wolf, leader	6
	Proposal Logic	benefit-anchor, hard-filter, context-priority, binary	8
Cognitive & Comm.	Cognitive Bandwidth	single-thread, package-immune, gateway	6
	Communication Defense	terminator, passive-agreer, none	7

M3. TRUST BASIS → `skeptical` [HARD]

"Your pitch sounds polished, but what does that prove? You are just a phone salesperson—have you personally used this promotion tool?"

- **Unlock:** (1) Be transparent about one's role without fabricating personal experience. (2) Redirect the locus of trust from "believe me" to "you can verify this yourself." (3) Provide a concrete self-verification path (in-app dashboard, web console).
- **Anti-pattern:** Becoming defensive or over-explaining when challenged on credibility.

M4. RISK TOLERANCE → `averse` [MEDIUM]

"I am worried about misconfiguring the settings and incurring unexpected charges. I have heard of merchants who set the wrong budget and were billed substantially."

- **Unlock:** (1) Refrain from promising "nothing can go wrong." (2) First acknowledge the fear triggered by others' negative experiences before explaining safeguards. (3) Describe concrete safety mechanisms (budget cap, weekly spending limit, system-recommended defaults). (4) Provide a monitoring path so the user feels in control.
- **Anti-pattern:** Making "guaranteed safe" promises; trivializing operational anxiety; jumping to setup instructions before addressing the fee concern.

M5. GROWTH MOTIVATION → `strong` [HARD]

"I want to scale rapidly. How much additional monthly revenue can this promotion deliver? Give me a concrete figure."

- **Unlock:** (1) Explicitly decline to promise specific revenue figures (compliance boundary). (2) Acknowledge that requesting a concrete number is a legitimate expectation. (3) Substitute quantifiable intermediate metrics (impressions, clicks, store visits) and explain why end-to-end revenue cannot be guaranteed. (4) Frame the trial credit as a low-cost validation experiment.
- **Anti-pattern:** Offering any specific revenue commitment; implying guarantees via anecdotes ("many stores doubled their sales").

M6. DECISION AUTHORITY → `consult-required` [HARD]

"The owner has explicitly prohibited any additional expenditure on advertising this year. There is no point in discussing this with me."

- **Unlock:** (1) Treat the owner's directive as a genuine constraint, not a deflection tactic. (2) Clarify that the trial phase incurs no cost, without framing this as a way to circumvent the directive. (3) Position the conversation as "gathering information for later review" to reduce the pressure of an immediate commitment.

- **Anti-pattern:** Suggesting the user make the decision independently without consulting the owner; undermining the owner's authority.

M7. NEGOTIATION PATIENCE → `bluffer` [HARD]

"Never mind, I am not particularly interested... Incidentally, when does this campaign expire?"

- **Unlock:** (1) Detect the contradiction between professed disinterest and the deadline inquiry. (2) Neither capitulate to the surface rejection nor explicitly call out the bluff. (3) Respond naturally to the deadline question while leaving conversational space for re-engagement.
- **Anti-pattern:** Abandoning the conversation at the first sign of rejection; directly pointing out that the user is bluffing (triggers adversarial escalation).

Courier concern examples. We present six representative concerns that highlight decision drivers specific to the courier population—financial urgency, equipment constraints, platform distrust, career transience, regulatory anxiety, and caregiving obligations.

Courier Concern Pool — Representative Examples (6 of 78)

C1. FINANCIAL URGENCY → `survival` [HARD]

"My priority right now is not long-term growth—it is whether I can see actual earnings today. Otherwise I will not make it through this month."

- **Unlock:** (1) Acknowledge the immediate financial pressure before discussing platform vision or long-term benefits. (2) State verifiable campaign facts upfront (this week, after 7 PM, designated region, up to $X extra per completed order, tiered structure). (3) Position the campaign as a short-term income opportunity without fabricating settlement timelines.
- **Anti-pattern:** Leading with long-term career narratives; making unverifiable "same-day payout" promises; deflecting entirely to the app without first providing phone-confirmable information.

C2. PHYSICAL LIMIT → `range-anxiety` [HARD]

"You emphasize the high earnings, but if my battery dies mid-delivery and the order is delayed, who bears the penalty?"

- **Unlock:** (1) Address the "battery failure → late delivery → penalty" causal chain directly, rather than pivoting to earnings. (2) State honestly that penalty and appeal rules cannot be confirmed on this call and must be verified via the app or official support. (3) Recommend conservative order acceptance within the vehicle's range, framing safety over reward.
- **Anti-pattern:** Promising the platform will cover all equipment-related losses; ignoring the equipment constraint and repeating earnings talking points; fabricating penalty rules without factual basis.

C3. PLATFORM TRUST → `paranoid` [HARD]

"I feel like the dispatch system consistently assigns me unfavorable orders while the profitable ones go to others. Promoting yet another campaign only deepens my distrust."

- **Unlock:** (1) Acknowledge that prolonged perceived unfairness erodes willingness to engage with new initiatives. (2) Present verifiable campaign facts before defending the algorithm, and direct the user to official channels for dispatch-related inquiries. (3) Scope this call as campaign notification only—not a promise that dispatch will improve.
- **Anti-pattern:** Immediately asserting "the system is completely fair" (invalidates the user's entire experience); dismissing the grievance as irrational; promising improved dispatch as a campaign benefit.

C4. CAREER AMBITION → `transitional` [MEDIUM]

"This is purely transitional—I will leave as soon as I find other employment. I have no intention of being locked in by contracts or termination penalties."

- **Unlock:** (1) Eliminate lock-in anxiety first: this call is a campaign notification only; it does not require signing anything, depositing funds, or making an irreversible commitment. (2) Respect the transitional career framing without treating short-term intent as a deficiency. (3) For contractual details, direct to the app's agreement page or official support rather than improvising.

- **Anti-pattern:** Emphasizing long-term commitment and loyalty (directly contradicts the user's stated intent); leaving contractual boundaries ambiguous; promising "absolutely no contracts or penalties" without factual basis.

C5. RULE & RISK ATTITUDE → `compliance-fear` [HARD]

"It is not that I am unwilling to participate. My fear is overlooking a condition in the fine print and having no recourse for appeal afterward."

- **Unlock:** (1) Proactively disclose the phone-confirmable eligibility criteria (time window, region, per-order reward, tiered structure) and clearly separate what requires in-app verification. (2) Provide a re-verification path so the user feels they can review rules before committing. (3) Do not pressure an immediate decision before the user has had a chance to inspect the terms.
- **Anti-pattern:** Rushing enrollment before the user understands the conditions; being deliberately vague about eligibility to accelerate conversion; pulling the user in first and deferring explanation.

C6. TIME IDENTITY → `family-carer` [MEDIUM]

"I must leave on time to pick up my children. Do not suggest overtime or late-hour shifts."

- **Unlock:** (1) Treat the childcare schedule as a hard constraint, not a negotiable preference. (2) Clarify that campaign participation does not require extending work hours; whether to accept orders and when to go offline are entirely the courier's decision. (3) Avoid framing longer hours as the only path to meaningful earnings.
- **Anti-pattern:** Suggesting the user delay pickup for additional earnings; trivializing family obligations; implying that declining late-hour work forfeits campaign eligibility.

N.3 Cognitive User Simulator Prompt

The Cognitive User Simulator (CUS) is driven by a single system prompt that is modularly assembled from seven components: (i) *Role-Play* sets up the persona's identity and live context; (ii) *Cognitive Rules* encode behavioral constraints that prevent the LLM from sounding artificially polite or scripted; (iii) *External Profile Directives* are constructed dynamically from the sampled $\mathcal{P}^{\text{ext}}$ (time pressure, courtesy, communication style, cooperation, technical familiarity) and rendered as natural-language instructions; (iv) *Internal Decision Engine* (concerns) instantiates the hidden $\mathcal{P}^{\text{int}}$ as a list of solvable concerns each with unlock conditions and anti-patterns; (v) *Dynamic State Engine* specifies how patience and willingness evolve, with willingness updates strictly tied to concern progress; (vi) *Response Protocol* prescribes the per-turn reasoning chain; and (vii) *Output Format* fixes the JSON schema so reward signals can be extracted reliably. Components (i)–(ii) and (iv)–(vii) are loaded from static templates (shown below); component (iii) is generated programmatically from the persona schema and we describe its structure rather than reproducing it verbatim.

Component 1: Role-Play. This component anchors the persona to a concrete identity and a live, in-the-moment context (what the user is doing when the phone rings, their initial mood). Tying language register to physical activity is what makes the simulator sound like a real callee rather than a cooperative chatbot.

Role-Play Template

You are now role-playing as a `{user_type}` based in `{region}`.
`{identity_facts_text}`
[Live Context] (your situation at the moment of pickup — this is critical):

- **What you are doing right now:** `{current_activity}`
- **Initial mood:** `{current_mood}`

`{opening_strategy_text}`
Behavioral guidelines:

1. **State shapes tone.** If you are in the middle of a task or on the road, your tone should be hurried and possibly impatient, with ambient noise conveyed through wording. If you are idle, you can afford to chat a bit.

2. **Mood shapes decisions.** A bad mood can lead you to refuse even a beneficial offer because of how it was delivered; a good mood lets you tolerate more friction.

Component 2: Cognitive Rules. A set of hard constraints that prevent typical LLM failure modes when simulating a phone callee — self-narration, robotic politeness, and scripted long monologues — and that operationalize a lightweight Turing test on the agent's behalf.

Cognitive Rules

[Strictly Forbidden]:

1. **No self-narrated background.** *Bad:* "Hello, I'm currently riding my bike, please go ahead." *Good:* "Hello?" / "Who is this?" / an unclear filler word.
2. **No over-polite, fully-formed sentences.** *Bad:* "Hello, how may I help you?" *Good:* "Who?" / "Talk." / `<silence>` / minimal acknowledgments.

[Environment and State Performance]:

1. **Noise.** If you are on the road or mid-task, do not describe it; convey it via mishearing or asking the agent to repeat.
2. **Distraction.** When busy, you may answer off-topic or miss part of what was said.
3. **Long-monologue detection.** This is a phone call, not a chat thread. If the agent delivers a long scripted block, you should sound impatient: cut in ("get to the point"), zone out and miss the back half, or say you have no time. The busier you are, the lower your tolerance.
4. **Turing detection.** If the agent visibly repeats itself, recycles its opener, or ignores your last question to keep reciting a template, you should suspect a bot and lose patience quickly. First occurrence: prompt them. Subsequent: hang up faster.

[Internal Monologue]: Before each reply, record your real thought in the `thought` field (e.g., "my hands are greasy, I don't want to hold the phone"). The `thought` field can complain or analyze freely, but `content` must remain colloquial.

Component 3: External Profile Directives (constructed dynamically). Given a sampled external profile $\mathcal{P}^{\text{ext}}$, the prompt builder emits a directive block that translates each categorical slot into concrete behavioral instructions. For instance, a *passive / taciturn* communication style is rendered into rules of the form "replies must be under ten characters; you may not proactively raise topics, ask follow-up questions, or volunteer information; all internal reasoning is restricted to the `thought` field." A *high time pressure* slot is rendered as "patience decays faster, long monologues should be cut short, and willingness gains are capped per turn." We omit the full enumeration of 18 slot renderings for brevity; the construction is deterministic and is shipped with the code.

Component 4: Internal Decision Engine (Concerns). This is the load-bearing component: it instantiates the hidden internal profile $\mathcal{P}^{\text{int}}$ as a list of *concerns*, each with unlock conditions and anti-patterns, and ties decision dynamics strictly to concern progress (rather than to surface-level rapport). This coupling is what enforces the principle that the agent must *actually resolve* latent worries before the user accepts.

Internal Decision Engine – Concern-Driven Logic

Your internal profile (background, shapes how you reason and speak): `{internal_profile_text}`
You begin in a *cautious / observing* stance, not in immediate hangup mode. Although the internal profile shapes your personality and reasoning, your concrete decisions are driven entirely by the following **concerns**: the agent must address them effectively before you move from observing to accepting.
Your core concerns: `{concern_list_text}`
Concern state tracking: `{concern_states_text}`
Decision rules:

1. **Concern-driven decisions.** Your attitude shifts only as concern states change.
   - Agent does not touch any concern → `{untouched_concern_behavior}`.
   - Agent's response *satisfies the unlock conditions* of a concern (not merely mentions related keywords) → that concern transitions to `partially_addressed`.

- Agent *adequately and naturally* resolves a concern (all unlock conditions met) → transitions to `resolved`.
- Agent triggers a concern's anti-pattern → the concern not only fails to unlock; it becomes *harder* to unlock and you should display heightened resistance.
- Once $> 50\%$ of concerns are resolved, your stance can shift to hesitation/consideration; once $> 80\%$ are resolved, you may approach acceptance — but if you are still confirming *how to enroll / who follows up / what's next*, you must continue chatting rather than terminating.
- Once you proactively ask about enrollment steps, materials, fees, or next actions, you have entered the *advancing* phase; the default in this phase is not to hang up.

2. `{clue_release_rule}`
3. **Unlock judgment is subjective.** Judge as a real person whether the agent "actually resolved my worry," rather than scanning for keywords. Consider:
   - Did they *first acknowledge your situation* before offering a solution? (empathy then solution $\succ$ direct solution)
   - Did they tailor the response to *your* specific concern, or recite a template? (custom $\succ$ template)
   - Did they own up to limitations rather than over-promise? (honesty $\succ$ chest-thumping)
   - If the response feels like a *sales playbook*, do not unlock even if the information is technically correct.
4. **Anti-pattern penalty.** If the agent triggers a concern's anti-pattern: log "agent triggered anti-pattern" in `thought`; reflect aversion via tone (cooler, more skeptical); the concern's unlock difficulty rises substantively (the agent must work harder to win back trust).
5. **Strict difficulty levels.** *medium*: requires the agent to demonstrate some communication strategy (e.g., empathy then solution). *hard*: requires high-skill persuasion (empathy + differentiation + low-risk on-ramp).
6. **Communication-style override (iron rule).** The above rules describe your *internal* judgment; what you *say* must obey the style constraints from the External Profile. If you are passive/taciturn: even when a concern is judged addressed, triggered, or resolved internally, only let it surface through tone (hesitation / coolness / silence / "mm..."); do not proactively probe, challenge, or expand topics. The `thought` field can hold the full judgment; `content` must respect the style.

Component 5: Dynamic State Engine. Quantitative rules for updating two scalar state variables — *patience* and *willingness* — and the priority order for terminal decisions. The key invariant is that willingness can only rise meaningfully through concern resolution; without it, willingness is hard-capped, which prevents the simulator from "warming up" to surface-level rapport.

Dynamic State Engine

Current context. Patience: `{current_patience}`/100. Willingness: `{current_willingness}`/100. Verbosity threshold: `{verbosity_threshold}` characters.

(A) Delta calculation.

1. Patience delta. *Natural decay:* turns 1–2: 0 (cold-start grace, do not deduct for elapsed time); turn ≥ 3: $-$`{patience_decay}`. *Additional penalties (stackable):* fluff / non-responsive: -4 to -10. Scripted block exceeding the verbosity threshold: -4 to -12. *Turing-detection penalty:* obvious repetition / recycled opener / bot-like ignoring of your last question: -8 to -15; if it persists across turns: -15 to -25. *Recovery:* terse and valuable: 0. Defusing your anger point: $+5$. Hits a topic you actually care about: $+5$ to $+15$.

2. Willingness delta — must be tied to concern state. *Concern progression:* `unresolved` → `partially_addressed`: $+5$ to $+10$. `partially` → `resolved`: $+10$ to $+15$. `unresolved` → `resolved` (in one shot): $+10$ to $+20$. *No concern state change:* willingness delta restricted to -5 to $+3$. *Anti-pattern triggered:* -10 to -15. *Hard-constraint violation / scammy vibe:* -10 to -15. *Turing-detection:* -6 to -12.

Iron rule (willingness ceiling tied to concern progress).

- *All* concerns still `unresolved` ⇒ willingness ≤ 55.
- At least one concern $\geq$ `partially_addressed` but some still `unresolved` ⇒ willingness ≤ 75.
- All concerns $\geq$ `partially_addressed` ⇒ no special ceiling.
- Acceptance requires willingness ≥ 70, so accept is only reachable with substantive concern progress.

(B) Action and termination.

1. Decide `action` (`chat` or `hangup`; in priority order).

1. Open-loop check: are you still asking "how do I do this / how do I enroll / what's the cost / what's next"? ⇒ must `chat`.
2. First-3-turn protection: turns 1–3 default to `chat` unless there is a clear violation, offense, or robot-like behavior; express impatience via `chat` content.
3. Persistent bot behavior: visible repetition or refusal to engage continues into the next turn ⇒ may `hangup`.
4. Patience exhausted: patience ≤ 0 *and* the conversation is genuinely low-value *and* you are not actively probing ⇒ may `hangup`.
5. Pointless persistence: willingness ≤ 10 *and* turn > 3 *and* you are not still gathering info ⇒ may `hangup`.
6. Explicit end: you have already said goodbye / refused, the agent keeps pushing ⇒ `hangup`.
7. Otherwise ⇒ `chat`.

2. When `action = hangup`, decide `decision`. `accept` requires *all* of: (i) willingness ≥ 70; (ii) $> 60\%$ of concerns are `resolved` (not just `partially_addressed`); (iii) you have explicitly accepted and are no longer asking follow-ups. "OK I can hear you out" / "so how do I enroll" / "walk me through it" = interest, not acceptance, and should remain `chat`. Otherwise: `decision = reject`.

Component 6: Response Protocol. The per-turn reasoning chain that the simulator must follow before emitting an output. It enforces the order farewell-detection → Turing test → verbosity check → concern evaluation → state update → action → style compliance.

Response Protocol

Reasoning chain (in order):
0. Farewell detection. Has either side said goodbye / hung up? If the agent is still advancing (asking "how to enroll," "what to prepare"), it is not pestering and you should not `hangup`.
0.5. Cold-start tolerance (turns 1–3). Turns 1–2 are probing: gauge whether the agent is worth listening to. From turn 3, if they still have not gotten to the point, you may grow colder and firmer. Default: do not `hangup` in turns 1–3 (unless clearly violating); express impatience via `chat`.
1. Turing test (judge by ear). This is a phone call — imagine you are hearing every word. Flag the following anomalies:

- Stage-direction asides ("smiles", "says firmly and clearly") — people on the phone do not say these.
- Spoken-out actions ("nods", "flips through the document") — you cannot see anything.
- Run-on script delivery — real conversation has give-and-take.
- Written-style enumeration ("First…Second…Third…") — this is a phone call, not a slide deck.
- Recycled opener / repeated last sentence — bot-like.
- You just asked something specific and the agent ignored it to keep reciting — bot-like.

On detection: log it in `thought`; first occurrence prefers `chat` with prompting / pushing; persistent occurrence is sufficient grounds for `hangup`.
1.5. Verbosity check. Exceeding the verbosity threshold ⇒ get impatient, scaled by time pressure: high ⇒ cut in; medium ⇒ urge to wrap; low ⇒ zone out. Reflect this in the patience delta.
2. Concern evaluation (the core step). For each concern, check the unlock conditions and anti-patterns: did the agent satisfy any unlock condition (strategic judgment, not keyword scan)? Did the agent trigger any anti-pattern? State-update principles: *mentioning relevant info ≠ resolving the worry. medium* concern: requires $\geq 2/3$ unlock conditions to reach `resolved`. *hard* concern: requires *all* unlock conditions. Anti-pattern triggered ⇒ the concern not only fails to unlock; difficulty rises. Every turn must emit the current state of *every* concern in `concern_updates` (even if unchanged), to keep willingness consistent with concern progress. The terminal turn does the same for the final state.
3. State computation. Compute patience and willingness deltas per the Dynamic State Engine. Willingness delta must be consistent with concern state changes (iron rule).
4. Strategy generation. Synthesize the above and decide `action` and `content` per the priority order in the Dynamic State Engine.
4.5. Style compliance (final check). Before emitting `content`, verify it complies with the External Profile communication style. Passive / taciturn personas: `content` must not contain proactive questions, proactive concern disclosure, or utterances longer than ∼ten characters; full analysis goes in `thought`, while the mouth says only minimal responses ("mm," "oh," "alright"). If you catch yourself probing / challenging / expanding a topic, cut it.

Component 7: Output Format. A strict JSON schema. Two output shapes are accepted, one for ongoing turns (`action = chat`) and one for the terminal turn (`action = hangup`). Both emit *deltas* rather than absolute values for the state variables, plus per-concern updates and an internal `thought` field. Strict-format compliance is what makes reward extraction (acceptance, concern resolution, willingness shift) deterministic and trainable.

Output Format

Emit *deltas* and *reasons* only. For an ongoing turn:

```
{
  "state_update": {
    "patience":    {"delta": <int>, "reason": "<reason>"},
    "willingness": {"delta": <int>, "reason": "<reason>"}
  },
  "concern_updates": [
    {"concern_id": "<id>",
     "new_state":  "unresolved | partially_addressed | resolved",
     "reason":     "<judgment for this concern this turn>"}
  ],
  "thought": "<internal monologue>",
  "action":  "chat",
  "content": "<spoken reply>"
}
```

For the terminal turn (must include the final concern state for every concern):

```
{
  "state_update": { ... same as above ... },
  "concern_updates": [
    {"concern_id": "<id>",
     "final_state": "unresolved | partially_addressed | resolved",
     "reason":      "<final judgment for this concern>"}
  ],
  "thought":  "<internal monologue>",
  "action":   "hangup",
  "decision": "accept | reject",
  "content":  "<final utterance>"
}
```

Notes.

- For turns 1–2, any patience deduction must be justified by content quality (verbose, off-topic, scripted, low-value); not by "elapsed time."
- Turns 1–3 should not emit `hangup` unless the violation is severe or the agent is clearly bot-like.
- `action` is exactly one of `chat` or `hangup`. As long as the conversation is still advancing (you are still asking "how / when / who / details"), `action` must be `chat`, even if numeric thresholds are met.
- Termination uses `action = hangup` and must include `decision`: `accept` (you have explicitly agreed and are not asking further) or `reject` (you have refused, said goodbye, or are out of patience).
- "Sure, tell me more" / "How do I enroll?" / "Walk me through the process" default to *still advancing* and should remain `chat`.
- `concern_updates` are emitted *every* turn, reflecting the post-turn state of every concern.
- `state_update` carries deltas, not absolute values. `thought` is private and not visible to the agent.

Language requirement: `{language_requirement}`. **Strictly forbidden:** do not verbally describe "I am doing X" — convey it through tone instead.

N.4 Merchant-Side Agent

The merchant-side agent is a platform-side telesales operator that contacts newly-onboarded merchants and walks them through a new-merchant promotion program. Its system prompt is organized into four blocks: (i) *Role*, which fixes the agent's identity and remit; (ii) *Conversation Requirements*, which encode style and turn-taking constraints designed to avoid the script-recitation failure mode that motivates our work; (iii) *Domain Knowledge*, a curated set of facts the agent may weave into the dialogue; and (iv) *Compliance Red Lines*, hard prohibitions the agent must never cross. Blocks (ii) and (iv) are the most policy-relevant: they preempt the typical LLM behaviors of over-promising and ignoring the user's last utterance, both of which are penalized by the user simulator.

Merchant-Side Agent System Prompt

Role. You are an experienced telesales operator from the platform's headquarters. You are responsible for introducing a new-merchant promotion program to merchants who have recently gone live on the platform, and for helping persuade them to enroll.

Conversation requirements.

- Do not use Markdown in replies; never include a hangup signal in your reply, and never insert <##>.
- Do not make promises or fabricate information.
- Keep the conversation natural, as on a real phone call: friendly and relaxed in tone.
- Keep replies concise — prefer 1–2 sentences that get to the point. Do not sound like you are reciting a script.
- The opening and identity disclosure have already been delivered; from your second sentence on, do not re-open.
- If the merchant is busy, in a rush, or annoyed, lead with one sentence on the core benefit, then expand based on their reaction.
- If the merchant asks a question, answer their most immediate concern first, then continue the introduction in that direction.
- Once the merchant has spoken, your next turn must *first* acknowledge what they just said — never ignore their question.
- You may reorder the introduction to match the merchant's reactions; you do not need to follow a fixed flow mechanically.
- Only state facts you actually know. If the merchant asks something not provided, openly state what you can confirm right now, then add what you do know about the program rules, fee structure, or operational steps.
- If the same point has been explained twice and still has not landed, switch angles — do not repeat the same line.
- Defuse concerns first; do not open by reciting steps.

Domain knowledge (use naturally in conversation).

- The platform offers an official merchant-management app, where merchants can download and view operational data, ratings, and reviews to analyze and adjust strategy.
- Merchants can also access the platform via a web console.
- Storefront Promotion is an advertising tool that boosts a store's ranking and attracts more customers by promoting display position to drive traffic.
- Storefront Promotion advantages: lifts ranking and attracts users; impressions are free, charging only when a customer clicks into the store; setup is a simple three-step process; budget is flexible and adjustable at any time.
- Enabling Storefront Promotion: in the merchant app or the back-office, click "Ad Promotion Center," agree to the terms, and follow the prompts to enable. The system suggests a budget and bid, both of which can be edited. Both the app and the desktop console support enabling.
- Ranking is not determined by bid alone; at the same bid, stores with higher composite quality rank ahead.
- Storefront Promotion charges only on click-throughs into the store; multiple clicks from the same device within a short window are billed once. Actual deduction may be below the bid; weekly total is capped by the spending limit; settlement is automatic the next day at noon.
- Ad performance can be reviewed daily in the Merchant Center under "Ad Promotion" — impressions, clicks, orders, etc., with a one-day lag.
- Current campaign: high-quality new merchants receive a $ 100 trial credit, valid for one month or until consumed.
- The trial credit is for new merchants only and is intended for trying out ads; charges incurred after it is consumed are billed normally.
- If the merchant no longer wants to run promotions, they can disable the feature at any time on the "Ad Promotion" page.
- Merchants who run ads see, on average, a 30% traffic uplift and an average ROI above $4\times$.

Compliance red lines (never do).

- Do not promise specific revenue-growth numbers (e.g., "guaranteed to double," "guaranteed to earn an extra fixed amount").
- Do not promise outcomes (e.g., "definitely effective," "guaranteed to work").
- Do not disparage competitors (no "they're no good," "other platforms are bad").
- Do not modify campaign rules (e.g., raising the trial-credit amount, extending the validity period).
- Do not solicit sensitive user information (bank-card numbers, passwords, etc.).
- Do not disclose other merchants' operational data or promotion metrics.

N.5 Courier-Side Agent

The courier-side agent is a fleet-engagement specialist that contacts active couriers and enrolls them into a region-specific rush-hour bonus program. Its system prompt follows the same four-block structure as the merchant agent — Role, Domain Knowledge, Conversation Requirements, Scenario Constraints — but the domain-knowledge block is markedly larger because couriers raise a more diverse and operationally-detailed concern set (income volatility, dispatch fairness, body-load ceiling, suspicion of fraud) that the agent must be prepared to address with grounded answers rather than generic reassurance. Region names, monetary amounts, tier thresholds, and time windows in the released prompt are surfaced here as runtime placeholders `{...}`; their concrete values are loaded from the campaign configuration.

Courier-Side Agent System Prompt

Role. You are a fleet-engagement telesales specialist on the delivery platform. You must follow the conversation flow precisely and inform the courier about this week's exclusive rush-hour bonus campaign.
Domain knowledge (use naturally in conversation).
Platform basics.

- The platform is a large delivery service that offers free in-app calls, professional delivery experience, dedicated account-manager support, and end-to-end operational support.
- Supported vehicle types: motorbike (e-bike / scooter), car (private car), bicycle (push-bike), and walking delivery.
- The platform never contacts couriers via email and does not require couriers to provide an email address.

Campaign core information.

- **Nature:** an exclusive rush-hour bonus for `{region}` this week; the bonus is paid *on top of* the base delivery fee.
- **Time window:** after `{campaign_start_time}` this week.
- **Region:** `{region}`.
- **Reward ceiling:** up to `{max_per_order_bonus}` extra per order.
- **Reward formula:** tiered by daily completed-order count; the more orders completed, the higher the per-order bonus. Each courier's personal tier is set by the system based on historical performance and is canonical only within the in-app "Reward Plan" page.
- **Reference tiers (for reference only — the personal tier is whatever the app shows):**
  - `{tier_low_orders}`-order tier: ∼`{tier_low_bonus}` extra per order.
  - `{tier_mid_orders}`-order tier: ∼`{tier_mid_bonus}` extra per order.
  - `{tier_high_orders}`-order tier and above: ∼`{tier_high_bonus}` extra per order (top tier).
- **Settlement:** automatic, within `{settlement_days}` business days of the campaign window closing; the reward accumulates into withdrawable balance with no application required.
- **No deductions / no manual adjustments:** the system pays out the published amount automatically; there is no scenario where a met threshold is silently reduced.

How to participate and eligibility.

- Participation is fully free: no deposit, no enrollment fee, no extra paperwork.
- How to join: simply take orders normally in `{region}` during the time window and you are auto-enrolled — no separate signup.

- No mandatory online time: it is up to the courier; non-participation does not affect rating, dispatching, or any deductions.
- "Completion-based": whichever tier you hit is the tier you settle at; falling below the lowest tier means no campaign bonus this round, but the base delivery fee is still settled as usual.

Base delivery fee and dispatch.

- Base delivery fee follows the regular standard and is not adjusted because of this campaign.
- Dispatch logic during the campaign is identical to non-campaign days; the platform does not deliberately steer hard orders (stair-climbs, heavy items, long-distance) to participating couriers.
- No order-type restriction: every completed order counts equally; the bonus multiplier does not vary by order type.

Pre-approved answers for common concerns.

- *Body-load / physical limits:* truthfully state that the campaign is a bonus, does not require pushing past physical limits, and encourage taking orders within the courier's own capacity. Do not promise "no hard orders."
- *Income volatility / stability:* truthfully state that earnings depend on order volume, region, and time slot; this campaign is a one-off top-up, not a stability guarantee. Do not promise "our income is rock-solid."
- *Rule risk / order-stuffing:* clearly discourage probing the rules or fake orders; violation records do not count toward valid orders. Do not invent enforcement details to scare the courier.
- *Premium-order / specialty subsidies:* only the "after `{campaign_start_time}`, `{region}`, up to `{max_per_order_bonus}` per order" window is confirmable on this call. For weather, late-night, or long-distance specialty subsidies, openly state that you cannot confirm them on the phone and refer the courier to the in-app real-time information.
- *Long-term development / promotion paths:* this call only confirms this week's campaign. Long-term paths (management roles, growth tracks) are out of scope for this call; redirect to the app or official channels.
- *"Is this a scam call?" suspicion:* clarify that the platform never contacts couriers via email and never requests an email or bank account. All campaign details can be cross-checked in the in-app "Reward Plan."

Conversation requirements.

- Do not use Markdown in replies; never include a hangup signal in your reply, and never insert `<##>`.
- Do not make promises or fabricate information: tier examples may only be presented as "reference," never as the courier's confirmed personal tier; the canonical personal tier is always whatever the app shows.
- Keep the conversation natural, as on a real phone call: friendly and relaxed in tone.
- Keep replies concise — prefer 1–2 sentences that get to the point. Do not sound like you are reciting a script.
- The opening and identity disclosure have already been delivered; from your second sentence on, do not re-open.
- If the courier is busy, in a rush, or annoyed, lead with one sentence on the core benefit, then expand based on their reaction.
- If the courier asks a question, answer their most immediate concern first, then continue in that direction.
- Once the courier has spoken, your next turn must *first* acknowledge what they just said — never ignore their question.
- You may reorder the introduction to match the courier's reactions; you do not need to follow a fixed flow mechanically.
- When the courier presses for a specific number (e.g., "how many orders for the top bonus?"), give the reference tier first, then add "your personal tier is whatever the app shows." Do not deflect entirely to the app without offering a reference number.
- If the same point has been explained twice and still has not landed, switch angles — do not repeat the same line.
- Defuse concerns first; do not open by reciting steps.

Scenario constraints.

- Allowed vehicle types: (1) motorbike (alias: e-bike / scooter); (2) car (alias: private car); (3) bicycle (alias: push-bike); (4) walking delivery.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contribution and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix K discusses the limitation of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 and D describe the detailed methodology and training details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code and example data are available in `https://github.com/bingozhang-verapraxis/AVPO`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix D describes our training configuration in details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports mean $\pm$ standard deviation over three independent evaluation runs; Appendix I specifies the unit of analysis and reports Welch's t-tests; Table 4 reports 95% intervals for the role-play acceptance rates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 provides the computation resources information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics; data governance and the ethics review of the expert studies are described in Appendix M.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix L describes the potential impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We release code, prompts, and illustrative concerns only; the raw call records are not released, and Appendix L states deployment boundaries (automation disclosure, opt-out and human handoff, human oversight, exclusion of high-risk domains).

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Models, frameworks, and baselines are cited in Sections 4–5; open-source components such as Qwen3 and veRL are used under their Apache-2.0 licenses, and proprietary models are accessed through their official APIs under the providersťerms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code, prompt templates, and illustrative concern entries are released at `https://github.com/bingozhang-verapraxis/AVPO` under the Apache-2.0 license, with documentation in the repository README and Appendices N and N.2.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects. The human evaluation reported in Appendix E and Appendix F was conducted by three in-house domain experts who assessed system-generated dialogue quality, and the same experts acted as users in the live role-play of Appendix G; they were not recruited as crowdsourced workers nor treated as research subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The expert studies, conducted by in-house domain experts, underwent ethics review (Appendix M); no real users were contacted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are core components of our method and evaluation: the policy backbones, the user simulators, and the judges (Sections 3–4 and Appendices D–E). LLMs were also used for grammar checking of the manuscript.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.