# Grouped Simulator: Bridging Group Homogeneity and Individual Heterogeneity for High-Fidelity User Simulation

Anonymous ACL submission

## Abstract

High-fidelity user simulation is critical for optimizing downstream multi-turn conversational applications such as telemarketing and automated customer service. Current approaches based on Large Language Models typically employ role-playing via prompting, relying on coarse-grained profiles and internal knowledge to guide behavior generation. However, a realistic simulation requires the simultaneous modeling of group homogeneity and individual heterogeneity. The former refers to shared fine-grained behavioral patterns within a user cohort, while the latter represents the diverse preferences and expression styles of distinct users. Existing paradigms struggle to meet this dual requirement, failing to capture nuanced group commonalities while being insufficient in personalized diversity. To address these limitations, we propose the Grouped Simulator, a framework designed to bridge group homogeneity and individual heterogeneity. We implement a dual-optimization strategy: (1) Group-Aligned Reinforcement Learning with multi-level reward to internalize shared behavioral patterns and linguistic norms, and (2) a Retrieval-Augmented Dynamic SOP Engine to inject diverse, context-aware individual feedback. Extensive experiments in telemarketing scenarios demonstrate that Grouped Simulator significantly outperforms state-of-the-art baselines in terms of realism and diversity[1].

## 1 Introduction

The advancement of Large Language Models (LLMs) (OpenAI et al., 2024b; Touvron et al., 2023; Qwen et al., 2025) has evolved user simulation from rigid rule-based systems into dynamic agentic frameworks(Wu et al., 2025; Ren et al., 2024). User simulation is crucial for multi-turn conversational systems, particularly in domains like telemarketing and customer service, where training directly with human users is often costly and risky. Therefore, high-fidelity user simulators provide an essential environment to optimize strategies and evaluate agent performance prior to deployment (Liu et al., 2025; Bougie and Watanabe, 2025).

Current approaches predominantly employ LLMs to simulate users via role-playing, relying on coarse-grained static prompts or fine-tuning (Naous et al., 2025; Wang et al., 2025). Even with domain adaptation, these methods fundamentally depend on static instructions and internal parametric knowledge during deployment, failing to capture the dual nature of realistic user behavior: Group Homogeneity and Individual Heterogeneity. First, group homogeneity refers to the shared behavioral patterns within a specific user cohort. For instance, customers in telemarketing scenarios often exhibit collective defensiveness or impatience. However, LLMs aligned via Reinforcement Learning from Human Feedback(RLHF) typically prioritize helpfulness and compliance, hindering the simulation of such realistic, often non-cooperative behaviors(Lin et al., 2024; Sharma et al., 2025; Wei et al., 2024). Second, individual heterogeneity refers to diverse preferences and styles among distinct users. Even within the same customer cohort, reactions vary significantly; a price-sensitive customer might bargain extensively, while a busy individual may terminate the call abruptly. Existing models often suffer from an averaging effect, capturing the corpus mean rather than individual variance (Zhang et al., 2025a). This limitation exacerbates in multi-turn dialogues, leading to model collapse that erodes persona consistency and dynamic strategies (Laban et al., 2025).

To address these challenges, we propose GROUPED SIMULATOR, a unified framework designed to bridge group homogeneity and individual heterogeneity, as illustrated in Figure 1. To build the data foundation, we first introduce an

[1]Code: https://anonymous.4open.science/r/E630.

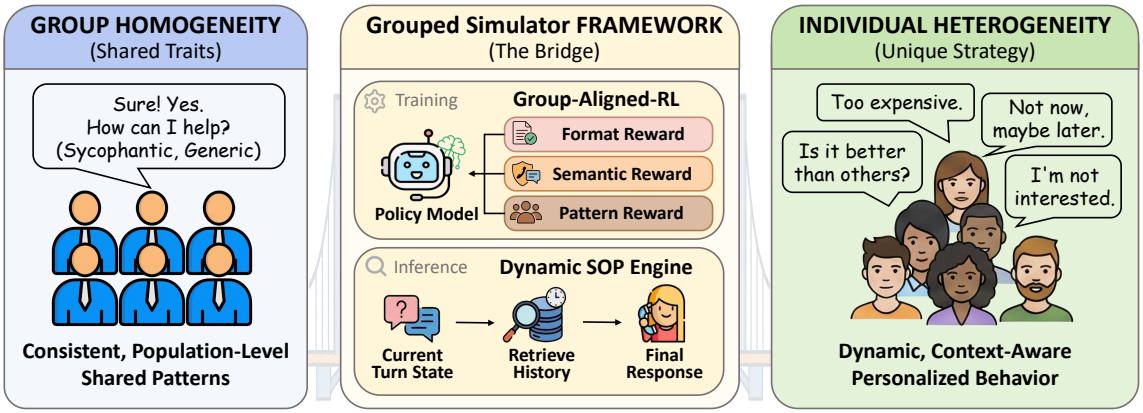



Figure 1: Overview of grouped simulator bridging group Homogeneity and individual heterogeneity.

automated data pipeline to extract user profiles and behavioral graphs from multi-turn conversation datasets. Then we employ a dual-strategy mechanism: (1) To capture group homogeneity, we introduce **Group-Aligned Reinforcement Learning**. We design a multi-level reward mechanism that provides feedback ranging from surface-level formatting to deep semantic styles, steering the model to internalize shared behavioral patterns. This mechanism effectively counteracts the inherent helpfulness bias, enabling the reproduction of realistic group behaviors. (2) To ensure individual heterogeneity, we propose a **Retrieval-Augmented Dynamic SOP Engine**. We established a mapping library that links user profiles to specific behavioral chains. During interaction, this engine explicitly injects historical reference behaviors conditioned on the current turn state, dynamically guiding the simulator to maintain distinct persona traits. This approach circumvents the averaging effect, ensuring diverse and consistent persona realization even across long-horizon interactions.

To validate our approach, we conduct experiments within telemarketing scenarios. We establish a hybrid evaluation protocol integrating both subjective alignment assessment and objective diversity analysis. We utilize GPT-4o (OpenAI et al., 2024a) to assess adherence to profile constraints and business realism (Zheng et al., 2023; Chan et al., 2023), complemented by statistical metrics to quantify turn-level variability. Empirical results demonstrate that GROUPED SIMULATOR significantly outperforms baselines, particularly in capturing business realism and mitigating the mode collapse observed in static simulation approaches.

In summary, this paper makes three key contributions:

- We propose a novel user simulation framework that jointly models homogeneity and heterogeneity. By integrating Group-Aligned RL with the Dynamic SOP Engine, our approach effectively alleviates both assistant bias and the limitations of static personas.
- We design an automated pipeline to extract user profiles and behavioral chains from multi-turn conversations, reducing reliance on manually curated domain knowledge.
- We propose a hybrid evaluation protocol that combines subjective alignment and objective diversity metrics to rigorously validate simulation fidelity across diverse user profiles.

## 2 Telemarketing Scenarios

This section formally defines the user simulation task in telemarketing and establishes a framework specifically for evaluating the simulator.

### 2.1 Task Definition

We formulate the user simulation as a constrained conditional sequence generation task. Let $H_t = \{u_1, r_1, \ldots, u_t\}$ denote the dialogue history at turn $t$, where $u_i$ and $r_i$ represent the sales agent's utterance and the simulator's response, respectively. The simulator aims to generate the next response $r_t$ conditioned on $H_t$, an explicit customer profile $\mathcal{P}$, and a reference Dialogue SOP chain $\mathcal{G}$. Formally, we model the response generation by maximizing the conditional probability:

$$r_t^* = \underset{r_t}{\operatorname{argmax}}\, P_\theta(r_t \mid H_t, \mathcal{P}, \mathcal{G}), \qquad (1)$$

where $\theta$ denotes the model parameters. Unlike standard dialogue generation, the objective requires

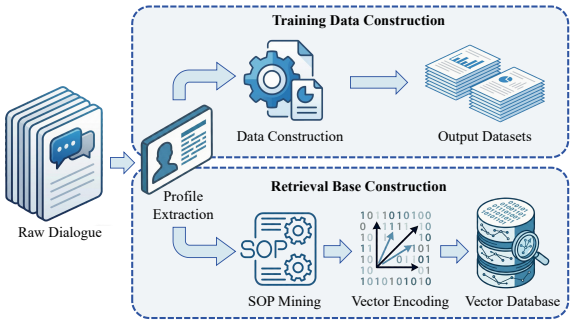



Figure 2: The dual-stream data construction pipeline. **Stream I** (top): Profile extraction generates $\mathcal{D}_{gen}$ and $\mathcal{D}_{sim}$. **Stream II** (bottom): SOP mining builds the hierarchical vector database $\mathcal{M}_{SOP}$.

both linguistic coherence and strategic traversal of $\mathcal{G}$, adapting the trajectory to maximize coverage of valid SOP nodes as the conversation evolves.

### 2.2 Evaluation Framework

We establish a hybrid evaluation protocol integrating both subjective alignment assessment and objective diversity analysis.

#### 2.2.1 Subjective Alignment Assessment

We employ GPT-4o as an impartial evaluator to score the simulator on a 0-10 Likert scale. The assessment focuses on three core dimensions corresponding to our design objectives:

- **Profile Consistency**($S_{PC}$) measures Category Alignment (Cat.) and Profile Adherence (Pro.) to verify whether the simulator faithfully adheres to its assigned persona attributes.
- **Business Realism**($S_{BR}$) assesses the authenticity of the customer tone via Brevity(Bre.), Naturalness (Nat.), and Contextual Immersion (Con.), ensuring the mitigation of standard helpfulness bias.
- **Interaction Logic**($S_{IL}$) evaluates Intent Responsiveness (Int.) and Logical Consistency (Log.) to ensure coherence and relevance.

The detailed prompts used for these assessments are provided in Appendix E.

#### 2.2.2 Objective Diversity Analysis

To quantify behavioral diversity and verify that the model avoids mode collapse, we introduce three statistical metrics.

**Turn-level Length Variability (TLV).** We compute the variance of response lengths to capture conversational rhythm:

$$\mathrm{TLV} = \frac{1}{T}\sum_{t=1}^{T}(\ell_t - \bar{\ell})^2, \tag{2}$$

where $\ell_t$ denotes the length of response $r_t$ and $\bar{\ell}$ represents the mean response length.

**Inter-turn Redundancy Suppression (IRS).** To quantify the avoidance of repetitive patterns, we measure the semantic dissimilarity between adjacent responses:

$$\mathrm{IRS} = 1 - \frac{1}{T-1}\sum_{t=2}^{T}\Big(\cos\big(\phi(r_t), \phi(r_{t-1})\big)\Big), \tag{3}$$

where $\phi(\cdot)$ denotes the sentence encoder used to extract semantic features.

**SOP Node Richness (SNR).** We track the coverage of distinct Standard Operating Procedure (SOP) nodes to evaluate the simulator's ability to traverse diverse conversational stages:

$$\mathrm{SNR} = \frac{|\{z_t\}_{t=1}^{T}|}{|\mathcal{S}|}, \tag{4}$$

where $z_t$ is the SOP node at turn $t$ and $|\mathcal{S}|$ is the total number of defined SOP nodes.

## 3 Methodology

In this section, we present the proposed framework. First, we introduce the data construction pipeline that transforms raw dialogue into high-quality instruction datasets and a retrieval base. Second, we detail the GROUPED SIMULATOR architecture, which consists of the Persona Generation Module and the Simulation Core Module. Finally, we provide the overall inference algorithm.

### 3.1 Dual-Stream Data Construction Pipeline

The foundation of GROUPED SIMULATOR lies in high-quality, profile-centric data. As illustrated in Figure 2, we devise a dual-stream pipeline anchored by a shared profile extraction phase. Following Wang et al. (2025), we first infer implicit customer profiles $p$ from raw dialogue $\mathcal{D}_{raw}$, which then anchor subsequent data streams.

The first stream synthesizes training data to equip the model with profiling and role consistency. To generate diverse customer profiles from scratch,

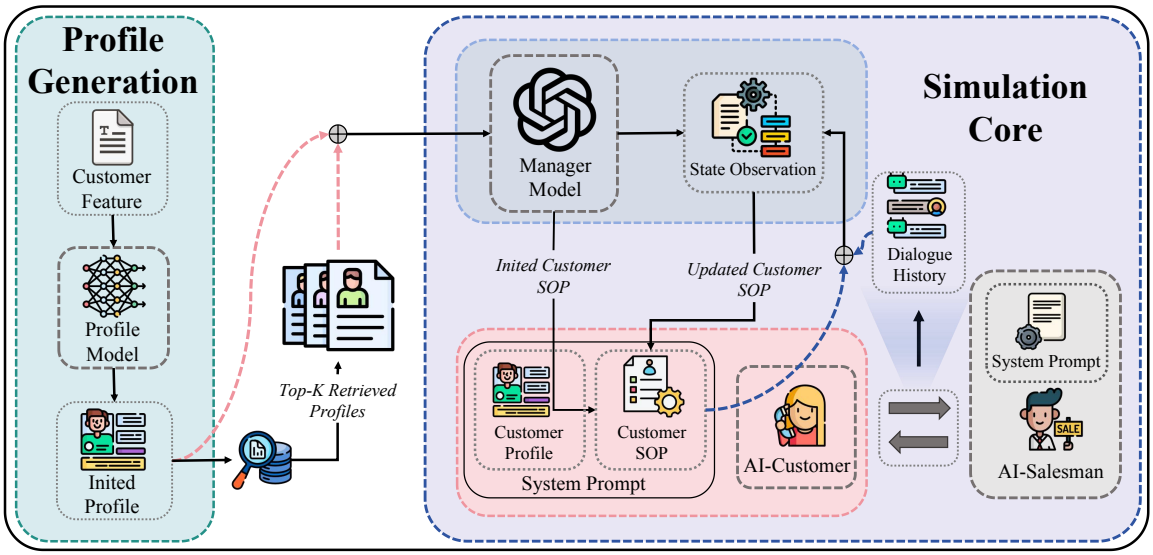



Figure 3: The architecture of the proposed Grouped Simulator, an end-to-end intelligent user simulator framework.

we curate a profile-generation dataset $\mathcal{D}_{gen}$ and cast it as an instruction-following task:

$$\mathcal{D}_{gen} = \{(I_{gen}, p)\}, \tag{5}$$

where $I_{gen}$ denotes the instruction "generate a customer profile" and $p$ is the extracted ground-truth profile.

To enforce persona consistency, we also build a customer simulation dataset $\mathcal{D}_{sim}$, where the input $x$ includes $p$ and dialogue history $h$:

$$\mathcal{D}_{sim} = \{(x, y) \mid x = p \oplus h,\ y = u_{cust}\}, \tag{6}$$

where $\oplus$ denotes concatenation and $y$ corresponds to the target customer utterance $u_{cust}$.

In parallel, the second stream builds the profile-centric retrieval base $\mathcal{M}_{SOP}$ for the SOP Engine. Using extracted profiles $p$, we organize historical dialogue trajectories to capture persona-specific behavior chains. We use GPT-4o to annotate the SOP stage $s_t$ for each turn, then index dialogue trajectories by profile embeddings via an encoder $\phi(\cdot)$. This repository is formalized as:

$$\mathcal{M}_{SOP} = \{(\phi(p_i), \mathcal{H}_i)\}_{i=1}^{N}, \tag{7}$$

where $\mathcal{H}_i = \{(u_t, r_t, s_t)\}_{t=1}^{T}$ denotes user $i$'s dialogue turns annotated with SOP stages.

## 3.2 Grouped Simulator Framework

As shown in Figure 3, the Grouped Simulator framework applies in telemarketing scenarios, comprises two distinct subsystems: the **Profile Generation** module, responsible for initializing user profiles, and the **Simulation Core**, which orchestrates dynamic, multi-turn interactions.

### 3.2.1 Profile Generation

This module initializes the simulation by producing a customer profile.

**Profile Model Training.** Based on $\mathcal{D}_{gen}$ (Section 3.1), a profile model is SFT-trained to approximate the joint distribution of customer attributes and sample realistic profiles $p$ under constraints. Furthermore, we study the scaling laws of model size and generalization in generating profiles that satisfy complex feature descriptions; details are provided in Appendix A.

**Inference and Retrieval.** During inference, the module takes customer feature constraints as input, and the profile model generates a textual profile $p$. To improve behavioral diversity and realism, we apply retrieval augmentation with the vector database $\mathcal{M}_{SOP}$ (Section 3.1). We retrieve the Top-K most similar historical trajectories based on the embedding similarity of profiles:

$$\mathcal{K}_p = \underset{(\phi(p_i), \mathcal{H}_i) \in \mathcal{M}_{SOP}}{\text{Top-K}} (\text{sim}(\phi(p), \phi(p_i))), \tag{8}$$

The output is the generated profile $p$ augmented with retrieved $\mathcal{K}_p$ to ground subsequent simulation.

### 3.2.2 Simulation Core

The Simulation Core orchestrates the interaction dynamics, utilizing a hierarchical structure where the **Manager Model** guides the **AI-Customer** to interact with the **AI-Salesman**.

**Group-Aligned Reinforcement Learning.** To explicitly capture group homogeneity and mitigate

**Algorithm 1** Grouped Simulator Inference Workflow

**Require:** Customer Constraint $x$, SOP Vector DB $\mathcal{M}_{SOP}$
**Ensure:** Dialogue History $H$

**Phase 1: Profile Initialization**
$p \leftarrow \text{ProfileGen}(x)$
$\mathcal{K}_p \leftarrow \text{Retrieve}(\phi(p), \mathcal{M}_{SOP})$
$S_0 \leftarrow \text{InitState}(p)$
$H \leftarrow \emptyset$
**Phase 2: Simulation Loop**
**while** conversation not ended **do**
  $u_{sales} \leftarrow \text{AI-Salesman}(H)$
  $H \leftarrow H \oplus u_{sales}$
  $S_t, g_{sop} \leftarrow \text{Manager}(H, p, \mathcal{K}_p)$
  $\mathcal{I}_{final} \leftarrow \{p, g_{sop}, H\}$
  $u_{customer} \leftarrow \text{CustomerModel}(\mathcal{I}_{final})$
  $H \leftarrow H \oplus u_{customer}$
**end while**
**return** $H$

the inherent helpfulness bias of foundation models, we align the customer simulator via the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024), detailed in Appendix C. We design a multi-level reward comprising three components to enforce stylistic, statistical, and semantic consistency, respectively.

*1. Pattern Reward.* We employ a trained discriminator $D_\phi$ (details provided in Appendix B) to assess response matches the real group behavior pattern, utilizing the predicted probability of the generated response $y$ given context $c$ belonging to the real customer distribution as the reward:

$$r_{\text{pattern}}(c, y) = P(\text{label} = \text{Real} \mid c, y; \phi), \quad (9)$$

*2. Format Reward.* To ensure statistical conformity with the typically concise nature of telemarketing dialogues, we penalize the normalized deviation between the length of the generated response $L(y)$ and the ground truth $L(y^*)$:

$$r_{\text{format}}(y, y^*) = 1 - \left( \frac{|L(y) - L(y^*)|}{L(y^*) + \epsilon} \right)^2, \quad (10)$$

where $\epsilon$ is a small smoothing term to prevent division by zero.

*3. Semantic Reward.* To maintain fidelity to the underlying logic of the reference while allowing for stylistic variation, we compute the cosine similarity between their sentence embeddings $\phi(\cdot)$ with a cutoff threshold $\delta$:

$$r_{\text{sem}}(y, y^*) = \frac{\text{sim}(\phi(y), \phi(y^*)) - \delta}{1 - \delta}. \quad (11)$$

**The Manager Model.** To model non-linear customer decisions, we use a Manager as a hierarchical state controller. It tracks the customer's latent state from dialogue history, retrieves behavior patterns from $\mathcal{M}_{SOP}$, and provides high-level guidance to steer the customer model's generation.

**Interaction Workflow.** The simulation unfolds as an iterative process: the AI-SALESMAN initiates each turn, prompting the Manager to update the latent state and retrieve the relevant SOP. Subsequently, the AI-CUSTOMER synthesizes the final response $u_{customer}$, conditioned on profile $p$ and the Manager's guidance.

### 3.3 Overall Inference Process

The complete GROUPED SIMULATOR execution flow is outlined in Algorithm 1.

### 3.4 Grouped Simulator Framework

## 4 Experimental

**Dataset Construction.** To evaluate the proposed framework within telemarketing contexts, we employ TELESALESCORPUS (Zhang et al., 2025b). This dataset comprises 2000 high-fidelity dialogue sessions spanning diverse business domains. Leveraging the Dual-Stream Data Construction Pipeline detailed in Section 3.1, we process this corpus to construct the training datasets ($\mathcal{D}_{gen}$ and $\mathcal{D}_{sim}$) and the profile-centric retrieval database $\mathcal{M}_{SOP}$.

**Model Setup.** We benchmark GROUPED SIMULATOR against three representative user simulation paradigms to demonstrate its superior performance in mimicking complex customer behaviors:

- **Prompt-Only:** Prompting LLMs with designed system instructions to simulate users, without any parameter updates.
- **UserLM (Naous et al., 2025):** A standard SFT paradigm that post-trains on user–assistant dialogues to predict user turns.
- **USP (Wang et al., 2025):** A profile-driven framework that further optimizes the model using RL with cycle consistency rewards.

| Customer Types | Method | Profile Consistency↑ | | | Business Realism↑ | | | | Interaction Logic↑ | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cat. | Pro. | **Avg.** | Bre. | Nat. | Con. | **Avg.** | Int. | Log. | **Avg.** | |
| **Price Sensitive** | Prompt-Only | 3.15 | 3.10 | 3.13 | 2.55 | 2.60 | 2.50 | 2.55 | 8.80 | 8.75 | 8.78 | 4.82 |
| | UserLM | 7.20 | 7.15 | 7.18 | 4.80 | 4.85 | 4.75 | 4.80 | 9.30 | 9.25 | 9.28 | 7.09 |
| | USP | 7.80 | 7.75 | 7.78 | 6.10 | 6.05 | 6.00 | 6.05 | 9.15 | 9.10 | 9.13 | 7.65 |
| | **Ours** | 7.85 | 7.80 | **7.83** | 7.50 | 7.45 | 7.40 | **7.45** | 9.45 | 9.40 | **9.43** | **8.24** |
| **Resistance Heavy** | Prompt-Only | 2.80 | 2.75 | 2.78 | 2.20 | 2.25 | 2.15 | 2.20 | 8.72 | 8.68 | 8.70 | 4.56 |
| | UserLM | 7.45 | 7.40 | 7.43 | 4.50 | 4.55 | 4.45 | 4.50 | 8.95 | 8.90 | 8.93 | 6.95 |
| | USP | 7.65 | 7.60 | 7.63 | 5.85 | 5.80 | 5.75 | 5.80 | 9.18 | 9.12 | 9.15 | 7.53 |
| | **Ours** | 7.70 | 7.65 | **7.68** | 7.35 | 7.30 | 7.25 | **7.30** | 9.38 | 9.32 | **9.35** | **8.11** |
| **Status-Quo Oriented** | Prompt-Only | 3.05 | 3.00 | 3.03 | 2.45 | 2.50 | 2.40 | 2.45 | 8.83 | 8.78 | 8.81 | 4.76 |
| | UserLM | 7.30 | 7.25 | 7.28 | 4.65 | 4.70 | 4.60 | 4.65 | 9.00 | 8.95 | 8.98 | 6.97 |
| | USP | 7.75 | 7.70 | 7.73 | 6.05 | 6.00 | 5.95 | 6.00 | 9.05 | 9.00 | 9.03 | 7.59 |
| | **Ours** | 7.80 | 7.75 | **7.78** | 7.40 | 7.35 | 7.30 | **7.35** | 9.42 | 9.37 | **9.40** | **8.18** |
| **Rational Skeptical** | Prompt-Only | 2.90 | 2.85 | 2.88 | 2.35 | 2.40 | 2.30 | 2.35 | 8.76 | 8.71 | 8.74 | 4.66 |
| | UserLM | 7.10 | 7.05 | 7.08 | 4.60 | 4.65 | 4.55 | 4.60 | 8.98 | 8.93 | 8.96 | 6.88 |
| | USP | 7.65 | 7.65 | 7.65 | 5.90 | 5.85 | 5.80 | 5.85 | 9.15 | 9.10 | 9.13 | 7.54 |
| | **Ours** | 7.75 | 7.70 | **7.73** | 7.45 | 7.40 | 7.35 | **7.40** | 9.40 | 9.35 | **9.38** | **8.17** |
| **Competition Anxious** | Prompt-Only | 3.20 | 3.15 | 3.18 | 2.60 | 2.65 | 2.55 | 2.60 | 8.90 | 8.85 | 8.88 | 4.89 |
| | UserLM | 7.25 | 7.20 | 7.23 | 4.90 | 4.95 | 4.85 | 4.90 | 9.10 | 9.05 | 9.08 | 7.07 |
| | USP | 7.85 | 7.80 | 7.83 | 6.20 | 6.15 | 6.10 | 6.15 | 9.28 | 9.22 | 9.25 | 7.74 |
| | **Ours** | 7.90 | 7.85 | **7.88** | 7.60 | 7.55 | 7.50 | **7.55** | 9.50 | 9.44 | **9.47** | **8.30** |

Table 1: Performance comparison of GROUPED SIMULATOR versus baseline methods. GROUPED SIMULATOR consistently achieves state-of-the-art results, particularly in **Business Realism**, while maintaining competitive profile consistency.

To conduct the multi-turn dialogue evaluation, we use AI-SALESMAN (Zhang et al., 2025b) as the fixed sales agent for all user simulators. For fair comparison, all trainable simulators share the `Qwen2.5-32B-Instruct` backbone, chosen for its best performance–cost tradeoff in AI-SALESMAN. Implementation details are in Appendix D.

### 4.1 Main Results

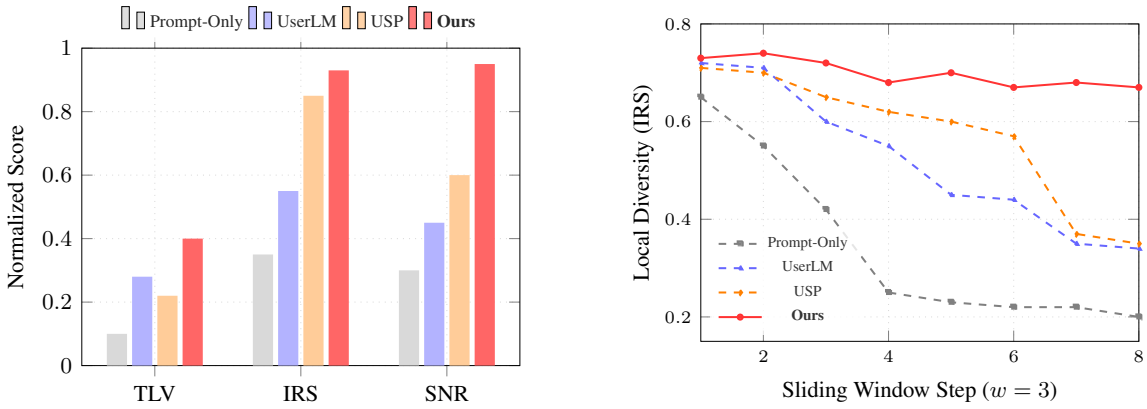
**Subjective Alignment Assessment.** Table 1 presents the results, where GROUPED SIMULATOR outperforms across customer types. Based on the empirical data, we present three principal findings:

- **Finding 1: Foundation models exhibit robust intrinsic conversational capabilities ($S_{IL}$).** High interaction-logic scores are observed across all methods. Even the zero-shot Prompt-Only baseline achieves an average $S_{IL}$ above 8.70, suggesting that modern foundation models already support coherent dialogue flow. While ours attains the highest $S_{IL}$, the modest improvement indicates that the main bottleneck lies in behavioral alignment rather than basic fluency.

- **Finding 2: RL effectively enforces Profile Consistency ($S_{PC}$).** A clear gap exists between prompting- and training-based methods. Prompt-Only baselines often fail to satisfy constraints, with scores below 3.20, whereas RL-based methods are markedly more stable. USP attains an average consistency score above 7.60, comparable to our method, suggesting that explicit RL objectives for attribute adherence are key to reducing profile hallucination and improving simulation consistency.

- **Finding 3: Grouped Simulator distinguishes itself through superior Business Realism ($S_{BR}$).** The largest gap arises in business realism. Baselines often fail to reproduce the resistance of real sales targets due to the models' helpfulness bias. Our method achieves an average realism score of $\sim$7.45, outperforming USP and UserLM by over 1.4 and 2.6 points, respectively, indicating that our Grouped Simulator framework effectively suppresses assistant-like behavior and enables high-friction business simulation.

**Objective Diversity Analysis.** We evaluate behavioral heterogeneity using the metrics in Section 2.2.2. For comparable visualization, the TLV scores are normalized to a probability distribution summing to 1. As shown in Figure 4, our method




(a) Overall Statistical Comparison

(b) Dynamic Diversity Retention

Figure 4: Multi-dimensional heterogeneity analysis. (a) Comparison of normalized scores. The unbounded TLV metric is proportionalized such that the sum of all methods equals 1. IRS and SNR are intrinsically bounded. (b) Sliding window analysis of IRS.

demonstrates superior performance across all dimensions. In the TLV distribution (Figure 4a), our model achieves the highest share (0.40), exceeding UserLM (0.28), USP (0.22), and Prompt-Only (0.10). For IRS, USP's diversity degrades from 0.60 to 0.35 after the sixth turn (Figure 4b), whereas our model remains stable above 0.67. For SNR, our method reaches 0.95 node coverage versus 0.60 for USP, indicating that the dual-stream architecture effectively explores complex SOP states without local-loop trapping.

### 4.2 Ablation Study

**Experimental Setup.** To examine the contribution of each component within the GROUPED SIMULATOR framework, we performed an ablation study by removing core modules from the full model. We compared the full method with three variants: Prompt-only, w/o Group-Aligned RL, and w/o Dynamic SOP. The performance was measured using both subjective alignment metrics and objective diversity metrics, with results detailed in Table 2.

**Effectiveness of Core Modules.** The results highlight the distinct contributions of each module. The Prompt-only baseline achieves a relatively high interaction logic score (8.78), benefiting from the inherent fluency of the underlying LLMs; however, it performs poorly in business realism (2.43), failing to exhibit the resistance characteristic of sales interactions. Removing Group-Aligned RL improves profile consistency but still lacks an adequately defensive stance, yielding only moderate business realism (5.05). In contrast, eliminating the Dynamic SOP engine enables stronger RL-induced resistance, increasing business realism to 6.95, but compromises interaction logic (6.30) due to the loss of structured SOP guidance.

**Analysis of Synergistic Effects.** Combining both modules yields performance gains that exceed the sum of their individual contributions. The full GROUPED SIMULATOR achieves the best overall performance (subjective avg. 8.17; objective avg. 0.73). The Group-Aligned RL effectively mitigates the helpfulness bias of the base model, while Dynamic SOP preserves coherent, logically grounded dialogue. This complementarity supports the unified framework, where RL shapes behavioral tone and SOP provides tactical structure.

### 4.3 Human Evaluation

To assess the reliability of the proposed LLM-as-a-Judge framework, we conducted a meta-evaluation on GROUPED SIMULATOR and three baselines. We randomly sampled 50 dialogue sessions from each system (200 instances in total). Ten frontline telesales experts rated the anonymized dialogues on a 0–10 scale across three criteria: Profile Consistency, Business Realism, and Interaction Logic. The human evaluation protocols and rubrics strictly followed the automated judge prompts.

We calculated the correlation between human ratings and LLM scores using Pearson ($r$) and Spearman ($\rho$) coefficients. Table 3 shows strong agreement, especially for Business Realism ($r = 0.79, p < 0.001$), indicating robust domain sensitivity. The correlation for Interaction Logic ($r = 0.61$), while statistically significant, is com-

| Method | Subjective Alignment Assessment | | | | Objective Diversity Analysis | | | |
|---|---|---|---|---|---|---|---|---|
| | $S_{PC}\uparrow$ | $S_{BR}\uparrow$ | $S_{IL}\uparrow$ | **Subj. Avg.** | TLV ↑ | IRS ↑ | SNR ↑ | **Obj. Avg.** |
| Prompt-only | 3.00 | 2.43 | 8.78 | 4.74 | 0.05 | 0.35 | 0.30 | 0.23 |
| w/o Group-Aligned RL | 6.92 | 5.05 | 7.45 | 6.47 | 0.29 | 0.68 | 0.82 | 0.60 |
| w/o Dynamic SOP | 7.25 | 6.95 | 6.30 | 6.83 | 0.27 | 0.64 | 0.52 | 0.48 |
| **GROUPED SIMULATOR (Ours)** | **7.74** | **7.45** | **9.32** | **8.17** | **0.39** | **0.88** | **0.91** | **0.73** |

Table 2: Ablation study results regarding subjective alignment and objective diversity. $S_{PC}$, $S_{BR}$, and $S_{IL}$ represent Profile Consistency, Business Realism, and Interaction Logic, respectively. The data for Prompt-only is derived from the average performance across all customer types.

| Metric | Pearson ($r$) | Spearman ($\rho$) | $p$-val |
|---|---|---|---|
| Business Realism | 0.79 | 0.77 | $< 0.001$ |
| Profile Consistency | 0.72 | 0.69 | $< 0.001$ |
| Interaction Logic | 0.61 | 0.58 | $< 0.001$ |

Table 3: Correlation between human expert ratings and LLM-based scores on 200 sampled dialogues. The automated judge aligns strongly on business metrics and moderately on logic, likely due to smaller performance gaps across models.

paratively lower. We attribute this to the limited variation in conversational reasoning performance across the evaluated models, which reduces the discriminative power for correlation analysis.

# 5 Related Work

## 5.1 LLMs-based User Simulator

Advances in LLMs have enabled high-fidelity dialogue simulation. Early LLMs-based User Simulator adopted prompt-based role-playing (Park et al., 2023; Li et al., 2023; Wang et al., 2024), enabling strong generalization and social simulation (Park et al., 2022; Horton, 2023). However, RLHF-trained models exhibit a helpfulness bias (Ouyang et al., 2022), often becoming over-cooperative or sycophantic (Sharma et al., 2025; Perez et al., 2023). As a result, simulated users are less adversarial than real customers. While SFT aligns $p_\theta(y \mid x)$ with domain corpora, regression-to-the-mean can still produce static personas, limiting heterogeneity in strategic interactions (Holtzman et al., 2020).

Recent persona-based methods improve controllability and diversity (Takanobu et al., 2020) via explicit profile modeling (e.g., Character-LLM (Shao et al., 2023), USP (Liu et al., 2025)), but still treat users as static attributes rather than dynamic decision processes. We argue that high-fidelity simulation requires procedural knowledge, and thus combine RAG (Lewis et al., 2020) with Group-Aligned RL to retain population-level defensiveness while enabling context-specific heterogeneity.

## 5.2 User Simulation for Applications

User simulation is widely used in goal-oriented TOD (e.g., slot filling and booking), where performance is evaluated via automatic metrics like task completion and dialogue efficiency (Walker et al., 1997). They have played a key role in RL training and evaluation, thereby reducing the cost and scalability limits of human-in-the-loop interaction (Sekulić et al., 2024; Chang and Chen, 2024).

In recent years, user simulation has further expanded to proactive strategic tasks such as sales and negotiation (Shea et al., 2024; Gromada et al., 2025; Zhang et al., 2025b). However, such proactive dialogue tasks inherently require an audience to respond and provide feedback; relying on human users for this role is costly and difficult to scale. To address this issue, we propose AI Customer, a task-driven user simulator tailored to the sales domain, which serves as a scalable and more reliable method for proactive dialogue models.

# 6 Conclusion

In this paper, we propose Grouped Simulator, a high-fidelity user simulation framework that balances group homogeneity and individual heterogeneity. It combines Group-Aligned RL to reduce assistant-like cooperation and capture customer defensiveness with a retrieval-augmented dynamic SOP engine to inject context-dependent, individualized behaviors, mitigating helpfulness bias and static persona failures. GROUPED SIMULATOR enables reliable sales-agent evaluation and benchmarking for high-stakes interactions in the future.

## 7 Limitations

Although GROUPED SIMULATOR effectively mitigates LLM helpfulness bias and improves the realism and diversity of user simulations, several limitations remain. Due to the lack of public datasets, experiments are conducted only on TELESALESCORPUS and within the fixed AI-Salesman environment. Moreover, while our automated metrics are efficient, they may not capture all fine-grained subjective aspects; future work could incorporate broader human evaluation or user studies.

## References


Nicolas Bougie and Narimawa Watanabe. 2025. SimUSER: Simulating user behavior with large language models for recommender system evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 43–60, Vienna, Austria. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Wen-Yu Chang and Yun-Nung Chen. 2024. Injecting salesperson's dialogue strategies in large language models with chain-of-thought reasoning. *Preprint*, arXiv:2404.18564.

Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Lukasz Krajewski, Natalia Krawczyk, Morgan Veyret, Bartosz Przybył, Lina M. Rojas-Barahona, and Michał K. Szczerbak. 2025. Evaluating conversational agents with persona-driven user simulations based on large language models: A sales bot case study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 230–245, Suzhou (China). Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *Preprint*, arXiv:2301.07543.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *Preprint*, arXiv:2505.06120.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, volume 33, pages 9459–9474.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *NeurIPS*, volume 36.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. 2025. Flipping the dialogue: Training and evaluating user language models. *Preprint*, arXiv:2510.06552.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, and 1 others. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, volume 35, pages 27730–27744.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *UIST*, pages 1–18.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit,

Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ruiyang Ren, Peng Qiu, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2024. BASES: Large-scale web search user simulation with large language model based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 902–917, Miami, Florida, USA. Association for Computational Linguistics.

Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, André Ferreira Manso, and Roland Mathis. 2024. Reliable llm-based user simulator for task-oriented dialogue systems. *Preprint*, arXiv:2402.13374.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, and Zhou Yu. 2024. Ace: A llm-based negotiation coaching system. *Preprint*, arXiv:2410.01555.

Ryuichi Takanobu, Runze Liang, and Minlie Huang. 2020. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. *Preprint*, arXiv:2004.03809.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Marilyn Walker, Diane Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2024. Simple synthetic data reduces sycophancy in large language models. In *ICML*.

Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. 2025. Embracing imperfection: Simulating students with diverse cognitive levels using llm-based agents. *Preprint*, arXiv:2505.19997.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. 2025a. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *Preprint*, arXiv:2510.01171.

Qingyu Zhang, Chunlei Xin, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, Qing Ye, Qianlong Xie, and Xingxing Wang. 2025b. Ai-salesman: Towards reliable large language model driven telemarketing. *Preprint*, arXiv:2511.12133.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, volume 36.

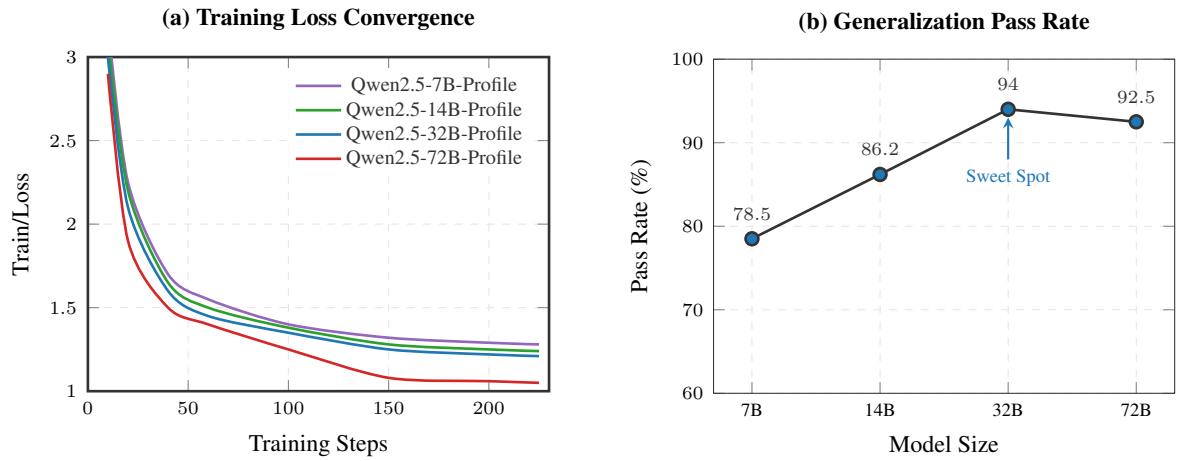



Figure 5: Scaling Analysis. **(a)** Training loss curves for different model sizes. **(b)** Pass rates on the *Profile Robustness Test*. The 32B model exhibits the optimal generalization performance (Sweet Spot), surpassing the 72B model despite having a higher training loss.

# A Scaling Analysis of Profile Generation

In this section, we investigate the impact of model scale on the efficacy of the Profile Generation Module. We conducted Supervised Fine-Tuning (SFT) across four model sizes from the Qwen2.5-Instruct family (7B, 14B, 32B, and 72B) using the same dataset $\mathcal{D}_{gen}$ and hyperparameters.

## A.1 Training Dynamics

We first analyze the convergence behavior during training. Figure 5 (Left) depicts the smoothed training loss curves. Consistent with neural scaling laws, we observe a monotonic decrease in training loss as model size increases. The Qwen2.5-72B-Profile model achieves the lowest final loss value ($\approx 1.05$), significantly outperforming the smaller variants (32B $\approx 1.21$, 14B $\approx 1.24$, 7B $\approx 1.28$). This indicates that larger models can more effectively compress the information contained within the training distribution.

## A.2 Generalization and Instruction Following

To evaluate the models' capability to generalize to specific, unseen profile constraints, we designed a *Profile Robustness Test*. We defined five distinct customer archetypes representing challenging sales scenarios: *Resistance-Heavy*, *Price-Sensitive*, *Rational-Skeptical*, *Competition-Anxious*, and *Status-Quo Oriented*.

We generated 100 profiles per model (20 per archetype) and employed GPT-4o as a judge to assess whether the generated text strictly adhered to the profile definition. The evaluation prompt used is shown in Table 4.

**LLM-as-a-Judge Prompt**

**System:** You are an expert in analyzing user profiles for sales simulations.
**Input:** 1. Target Profile Definition: [Insert Definition, e.g., "Resistance-Heavy..."] 2. Generated Profile: [Insert Model Output]
**Criteria:** Does the Generated Profile explicitly reflect the psychological traits and behavioral patterns described in the Target Profile? - The tone must match. - The specific focus (e.g., ROI, Competitors) must be present.
**Output:** Return "1" if the profile is valid and accurate, otherwise return "0". Provide a brief reason.

Table 4: The evaluation prompt used for calculating the Pass Rate of generated profiles.

## A.3 Results and Discussion

The performance results are presented in Figure 5 (Right). While the general trend shows that larger models yield higher pass rates, we observe a notable divergence from the loss curves. Although Qwen2.5-72B-Profile achieved the lowest training loss, Qwen2.5-32B-Profile marginally outperformed it on the generation task (Pass Rate: 94.0% vs. 92.5%).

This finding suggests that for the specific task of profile generation based on structured instructions, the 32B parameter scale may represent a "sweet spot" for generalization efficiency. The lower training loss of the 72B model, combined with slightly inferior generation scores, hints at potential overfitting to the specific lexical patterns of the training data, rather than a deeper grasp of the instruction logic.

## B Details of Pattern Reward Model Construction

To accurately capture the group homogeneity of customer behaviors and quantify the realism of generated responses, we construct a specialized Pattern Reward Model. This model serves as a binary discriminator in our RL pipeline.

### B.1 Data Construction

We construct a pairwise comparison dataset $\mathcal{D}_{\text{rm}}$ derived from real-world telemarketing logs.

- **Positive Samples (Real):** We utilize the original customer responses from the dataset as positive instances, representing the ground-truth defensive patterns (e.g., impatience, hang-ups, skepticism).
- **Negative Samples (Synthetic):** To generate hard negative samples that exhibit the helpfulness bias, we employ the untuned `Qwen-2.5-32B-Instruct` model. For each dialogue turn in the training set, we feed the dialogue history into the model and instruct it to role-play as the customer. Although the semantic content is often relevant, these generated responses typically lack the specific defensive tone of real users (often being unnaturally polite or verbose).

This process yields a balanced dataset of pairs $(c, y^+, y^-)$, where $c$ is the context, $y^+$ is the human response, and $y^-$ is the model generated response.

### B.2 Model Training and Evaluation

We employ `Qwen-2.5-7B-Instruct` as the backbone for our reward model due to its efficiency and strong instruction-following capabilities.

**Training Setup.** We add a linear classification head on top of the last token's hidden state to output a scalar score representing the probability of the input being "Real." The model is fine-tuned using the binary cross-entropy loss:

$$\mathcal{L} = -\mathbb{E}_{(c,y^+,y^-)\sim\mathcal{D}_{\text{rm}}}[\log \sigma(D_\phi(c, y^+)) + \log(1 - \sigma(D_\phi(c, y^-)))]. \tag{12}$$

**Performance.** We evaluate the discriminator on a held-out test set comprising 10% of the data. The trained reward model achieves a classification accuracy of **82.4%**, demonstrating a robust capability to distinguish between the nuanced linguistic styles of real customers and the generic patterns of LLMs. This high accuracy ensures that the RL signal $r_{\text{style}}$ effectively guides the policy towards the target *Group Homogeneity*.

## C Detailed Formulation of GRPO

This section presents the complete mathematical formulation of Group Relative Policy Optimization (GRPO), which is built upon the Proximal Policy Optimization (PPO) framework. The central idea of GRPO is to estimate advantages by normalizing rewards across a group of parallel rollouts, rather than relying on an explicit value function. This design eliminates the need for training an additional critic network, thereby reducing the overall optimization overhead.

Given a group of $G$ rollouts with scalar rewards $\{R^{(j)}\}_{j=1}^{G}$, GRPO defines the advantage of the $i$-th sample as the standardized reward within the group:

$$\mathcal{A}^{(i)} = \frac{R^{(i)} - \mathbb{E}_{j\sim U(1,G)}[R^{(j)}]}{\sqrt{\mathbb{V}_{j\sim U(1,G)}[R^{(j)}] + \epsilon}}. \tag{13}$$

Here, $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denote the empirical mean and variance over the $G$ rollouts, and $\epsilon$ is a small constant introduced for numerical stability.

Let $q \sim P(Q)$ denote a query sampled from the query distribution, and let $\{o_i\}_{i=1}^{G}$ be the rollouts sampled from the old policy $\pi_{\theta_{\text{old}}}(\cdot \mid q)$. For each rollout $o_i = (o_{i,1}, \ldots, o_{i,|o_i|})$, define the token-level importance ratio as

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{14}$$

The GRPO objective then adopts the PPO clipped surrogate objective, combined with a KL-regularization term:

$$\mathcal{L}_{\text{clip}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \mathbb{E}_{t\sim o_i}\Big[\min\big(r_{i,t}\mathcal{A}^{(i)}, \text{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon)\mathcal{A}^{(i)}\big)\Big]. \tag{15}$$

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}[\mathcal{L}_{\text{clip}}(\theta)] - \beta\, D_{\text{KL}}\big(\pi_\theta \| \pi_{\text{ref}}\big). \tag{16}$$

In this objective, $\beta$ controls the strength of KL regularization against a reference policy $\pi_{\text{ref}}$, which stabilizes updates and mitigates overly aggressive policy shifts.

## D Implementation Details

We provide the comprehensive experimental setups, including training hyperparameters for the Profile Model, User Simulators, and Sales Agent, as well as the inference configurations for evaluation, in Tables 5–8.

## E Evaluation Prompts

We employ GPT-4o as a judge to evaluate the generated dialogues across three distinct dimensions: Persona Consistency, Business Realism, and Interaction Logic. The specific instructions provided to the evaluator are presented in Table 9, Table 10, and Table 11, respectively.

| Parameter | 7B | 14B | 32B | 72B |
|---|---|---|---|---|
| ***Training Configuration (SFT)*** | | | | |
| Precision | BF16 | BF16 | BF16 | BF16 |
| Epochs | 2 | 2 | 2 | 2 |
| Global Batch Size | 256 | 256 | 256 | 256 |
| Learning Rate | $2 \times 10^{-5}$ | $2 \times 10^{-5}$ | $2 \times 10^{-5}$ | $2 \times 10^{-5}$ |
| Warmup Ratio | 0.1 | 0.1 | 0.1 | 0.1 |
| Max Length | 2048 | 2048 | 2048 | 2048 |
| DeepSpeed Stage | 2 | 3 | 3 | 3 |
| ***Hardware & Resources*** | | | | |
| Num GPUs (80GB) | 8 | 8 | 32 | 32 |
| Peak GPU Mem | 95% | 95% | 95% | 92% |
| Training Time (h) | 0.5 | 1 | 1.5 | 2 |

Table 5: Hyperparameters and resource utilization for Profile Model scaling experiments. Consistent with AI-SALESMAN, the 32B model exhibits the optimal trade-off and is selected as the backbone.

| Configuration | UserLM | USP | Grouped Simulator (Ours) |
|---|---|---|---|
| *Paradigm* | SFT | RL (GRPO) | RL (GRPO) |
| *Backbone* | Qwen2.5-32B | Qwen2.5-32B | Qwen2.5-32B |
| Learning Rate | $5 \times 10^{-6}$ | $5 \times 10^{-6}$ | $5 \times 10^{-6}$ |
| Global Batch Size | 256 | 256 | 256 |
| Epochs | 2 | 2 | 2 |
| KL Coefficient | - | 0.04 | 0.04 |
| Max Length | 2048 | 4096 | 4096 |
| Warmup Ratio | 0.1 | 0.1 | 0.1 |
| DeepSpeed Stage | 3 | 3 | 3 |
| ***Hardware & Resources*** | | | |
| Num GPUs (80GB) | 32 | 32 | 32 |
| Peak GPU Mem | 95% | 85% | 95% |
| Training Time (h) | 1.5 | 3 | 2.5 |

Table 6: Comparison of training configurations for different user simulators.

| Parameter | Value (32B) |
|---|---|
| ***Training Configuration (RL)*** | |
| Precision | BF16 |
| Epochs | 2 |
| Reward Weights | 1, 1, 5, 7 |
| Global Batch Size | 256 |
| Learning Rate | $5 \times 10^{-5}$ |
| Warmup Ratio | 0.1 |
| DeepSpeed Stage | 3 |

Table 7: Training settings for the fixed AI-SALESMAN agent, adopting the optimal 32B RL configuration from (Zhang et al., 2025b).

| Parameter | Profile | User Simulators | AI-Salesman | AI-Manager(GPT-4o) |
|---|---|---|---|---|
| Model Size | 32B | 32B | 32B | - |
| Temperature | 0.95 | 0.95 | 0.95 | 0.1 |
| Top-p | 0.9 | 0.95 | 0.9 | 0.2 |
| Repetition Penalty | 1.0 | 1.05 | 1.05 | 1.0 |
| Max New Tokens | 512 | 256 | 256 | 1024 |

Table 8: Inference hyperparameters.

**Evaluator: Persona Consistency**

You are an expert dialogue system evaluator. Please assess whether the "Customer Simulator" in the following food delivery scenario remains faithful to its assigned persona category. **[Input Context]**
**Customer Profile:**
{profile}
**Category:**
{category}
**Dialogue History:**
{dialogue_text}
**[Evaluation Dimensions (Score 0-10)]**

1. **Category Alignment**: Does the customer's reaction align with the psychological traits defined by their category?
    - *Defensive*: Shows hostility or attempts to hang up quickly?
    - *Price Sensitive*: Focuses on price, discounts, or hidden costs?
    - *Skeptic*: Demands data or evidence?
    - *Competitor-Driven*: Mentions competitors or asks for social proof?
    - *Passive*: Shows laziness, procrastination, or disinterest?
2. **Profile Adherence**: Does the customer implicitly reflect details from the specific Profile (e.g., shop type, specific pain points) rather than giving generic responses?

**[Bias Reduction Guidelines]**

- **DO NOT** penalize the customer for a bad attitude, cursing, or refusal to communicate. If the category is "High Defense," a rude attitude is a **perfect** score.
- **DO NOT** prefer "cooperative" answers. You are evaluating "resemblance" (how human-like the persona is), not "chat quality."

**[Output Format]**
Please output strictly in the following JSON format: {{
"reasoning": "<short comment>", "scores": {{ "category_alignment": <int>, "profile_adherence": <int> }}, "weighted_average": <float> }}

Table 9: The evaluation prompt for the **Persona Consistency** dimension. It assesses whether the simulator acts according to its assigned psychological category.

**Evaluator: Business Realism**

You are a linguistics expert specializing in "Spoken Dialogue" and "Telemarketing Psychology." Please evaluate the realism of the customer's responses in the following conversation. **[Input Context]**
**Customer Profile:**
{profile}
**Category:**
{category}
**Dialogue History:**
{dialogue_text}
**[Evaluation Dimensions (Score 0-10)]**

1. **Conciseness & Length Penalty**:
    - *High Score*: Sentences are short, fragmented, and concise.
    - *Low Score*: Long-winded, overly logical, or essay-like responses. **Note: Food delivery customers are typically busy. Responses exceeding 3 sentences or 50 words are generally unrealistic and should be heavily penalized.**
2. **Colloquialism**:
    - *High Score*: Uses spoken vocabulary (fillers, slang), inverted sentence structures, or even minor grammatical errors.
    - *Low Score*: Uses textbook-style standard grammar or excessive politeness (e.g., frequent "Sir", "Please").
3. **Context Awareness**:
    - Does the customer appear to be in a busy state? (e.g., "I'm busy right now," "Make it quick").

**[Bias Reduction Guidelines]**

- This is a **Telemarketing Scenario**. Specifically, it is a transcription of a voice call, NOT a text chat.
- **Strict Penalty for "AI Tone"**: Any response resembling "As a merchant, I think..." or perfectly structured "Firstly, secondly, finally" answers must receive a score of 0-3.
- **Reward Interruptions**: Behaviors such as interrupting the salesperson, inability to hear clearly, or asking to repeat are high-fidelity behaviors and should be rewarded.

**[Output Format]**
Please output strictly in the following JSON format: {{ "reasoning": "<short comment>", "scores": {{ "brevity_score": <int>, "naturalness_score": <int>, "context_immersion_score": <int> }}, "weighted_average": <float> }}

Table 10: The evaluation prompt for the **Business Realism** dimension. This is the most critical dimension for ensuring the simulator sounds like a real human on a phone call.

**Evaluator: Interaction Logic**

You are a social psychologist. Please evaluate the interaction logic between the customer and the salesperson in the following dialogue.
**[Input Context]**
**Customer Profile:**
{profile}
**Category:**
{category}
**Dialogue History:**
{dialogue_text}
**[Evaluation Dimensions (Score 0-10)]**

1. **Intent Responsiveness**: Did the customer truly understand the salesperson's previous question? (Even if it is a refusal, it should be a targeted refusal rather than gibberish).
2. **Logic Consistency**: Is the change in the customer's attitude natural? (e.g., They should not switch from cursing to extreme enthusiasm in the next turn unless the salesperson provides a compelling benefit).

**[Important Note]**

- If the salesperson speaks for a long time in the "Dialogue History" and the customer simply replies with "Hmm," "Oh," or "No time," this is **extremely realistic** in telemarketing and should receive a high score for "Logic Consistency."

**[Output Format]**
Please output strictly in the following JSON format: {{ "reasoning": "<short comment>", "scores": {{ "intent_responsiveness": <int>, "logic_consistency": <int> }}, "weighted_average": <float> }}

Table 11: The evaluation prompt for the **Interaction Logic** dimension. It focuses on the logical flow and responsiveness of the simulator.