TRACER: TRAJECTORY-ALIGNED LEARNING FOR MULTI-TURN USER SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Faithful user simulation is fundamental to building, evaluating, and improving interactive AI at scale. Yet current simulators often produce plausible individual responses without reproducing the intent evolution and outcomes observed in real interactions. We propose TRACER, a multi-turn user simulator that models evolving user intent and aligns simulated trajectories with real ones. TRACER is trained in two stages: supervised fine-tuning on real user dialogues, followed by multi-turn reinforcement learning. The RL stage combines hierarchical outcome- and trajectory-level rewards with deviation-aware advantage modulation, jointly addressing reward sparsity and credit assignment challenges in long dialogues. On real customer-service sessions organized into reference cohorts, TRACER-7B surpasses the strongest baseline by 11.4 conversion F1 points, while outperforming all baselines on group-level conversion-rate error and semantic trajectory distance and generalizing to out-of-distribution scenarios. In human Turing tests, annotators identified TRACER conversations at near-chance accuracy. Building on this simulator, we further introduce the Dynamic Marketing Benchmark, which jointly evaluates persuasion and response quality via simulated interactions, revealing that higher response quality does not necessarily correspond to higher conversion rates.

1 INTRODUCTION

With the widespread application of large language models (LLMs) in dialogue systems, scalable user feedback has become increasingly important for model evaluation and improvement. User simulators provide such feedback through interactions with dialogue agents (Hu et al., 2023; Abbasiantaeb et al., 2024; Qiu & Lan, 2024; Dou et al., 2025), supporting offline evaluation and reinforcement learning (Wu et al., 2025; Wang et al., 2025b; Qian et al., 2025b). Their usefulness, however, depends not only on linguistically natural responses but also on whether simulated users make decisions consistent with real human behavior across the interaction. Multi-turn simulation must therefore account for how users express themselves and how their behavior unfolds over complete interactions.

However, existing user simulators still suffer from two fundamental limitations. *(L1) Alignment is limited to linguistic style, not decision-making.* Mainstream methods condition LLMs on static user profiles to elicit persona-consistent responses (Wang et al., 2025a; Shi et al., 2025; Ye et al., 2025; Kim & Yang, 2025), aligning simulators with real users only at the surface linguistic level while leaving the underlying decision-making process unaligned. *(L2) Optimization is single-turn, not trajectory-level.* Recent methods adopt optimization paradigms tailored to turn-level responses (Wang et al., 2025c; Liao et al., 2025; Zhang et al., 2026; Wu et al., 2026; Dou et al., 2025), and thus inherently struggle with reward sparsity and credit assignment in multi-turn interactions, failing to capture state evolution and decision coherence. Consequently, simulated feedback often exhibits behavioral instability and inconsistent decision-making, limiting its utility for high-fidelity evaluation and training.

To address these limitations, we propose **TR**ajectory-**A**ligned **C**redit Assignment for Us**ER** Simulation (**TRACER**), a multi-turn user simulator trained on real dialogues. TRACER combines linguistic grounding with behavioral alignment, using intent states inferred from observed dialogue as proxies for users' decision tendencies. In the first stage, supervised fine-tuning learns user-like responses and initializes structured intent prediction. In the second stage, the simulator interacts with a fixed assistant and learns from rewards measuring terminal agreement and trajectory alignment against

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

paired human logs. Trajectory-level rewards are further decomposed to credit individual turns, using deviation magnitude and persistence to calibrate per-turn updates without overriding the session-level signal.

To evaluate behavioral fidelity, we organize 3,866 real customer-service sessions into 762 reference groups based on similar user conditions. Each group provides multiple observed trajectories, allowing evaluation to account for variation among comparable users and reducing dependence on a single reference interaction. We compare simulated and real behavior within these groups along three dimensions: session-level behavior, dialogue structure, and semantic consistency. We also assess generalization to unseen scenarios and perceived human-likeness. In a Turing study, annotators identified TRACER conversations at near-chance accuracy, suggesting high perceived naturalness.

Building on TRACER, we further introduce the Dynamic Marketing Benchmark (DM-Bench), which evaluates LLMs as conversational marketing agents through interactions with simulated users. DM-Bench jointly measures response quality and conversion outcomes, providing a behavioral perspective on agent performance. Within this simulated environment, models with higher response quality do not necessarily achieve higher conversion rates, motivating evaluation of both dimensions.

We summarize our contributions as follows: 1) We propose TRACER, a two-stage user simulator that combines outcome and trajectory rewards with deviation-aware turn-level advantage modulation, which uses the magnitude and persistence of trajectory deviations to adjust per-turn updates while preserving session-level advantage direction. 2) We construct a multi-reference evaluation set with 762 groups and 3,866 real sessions, supporting assessment of session behavior, dialogue structure, and semantic consistency. 3) We develop DM-Bench to jointly evaluate response quality and conversion outcomes in simulated marketing interactions, revealing a misalignment between the two.

2 Related Work

LLM-based User Simulator. Recent LLM-based user simulators have evolved from prompt-based role-playing toward deeper cognitive modeling. Early works often rely on explicit personas or static intent predictions, capturing surface linguistic patterns or individual responses but struggling to model dynamic intent evolution across multi-turn interactions (Park et al., 2024; Wang et al., 2025a; Kim & Yang, 2025; Kolluri et al., 2025). Recent advances incorporate explicit reasoning and state modeling. HUMANLM (Wu et al., 2026) uses reinforcement learning to align latent states and generated responses with observed responses, while Lu et al. (2025) use reasoning-augmented supervised fine-tuning to predict the next user action from interaction histories. UserLM-R1 (Zhang et al., 2026) improves reasoning and strategic behavior through multi-reward reinforcement learning over generated rationales and responses. Although these methods leverage multi-turn context, their optimization focuses on individual responses or next-step actions rather than aligning complete rollouts with human reference trajectories. Overall, existing simulators fall short in capturing how continuous behavioral evolution shapes final dialogue outcomes.

Applications of user simulator. User simulators are heavily utilized in dialogue system research for both training and evaluation, alleviating the reliance on costly human annotation. **Training:** Simulators provide scalable interactive environments, either by generating large-scale dialogue corpora (Li et al., 2023) or by supplying multi-turn dynamic feedback and reward signals in RL loops (Wang et al., 2025b; Qian et al., 2025b; Yang et al., 2026; Wu et al., 2025). **Evaluation:** Simulators overcome the limits of static benchmarks by enabling diverse and realistic interaction assessments (Zhang et al., 2025; Qian et al., 2025a). However, the reliability of these simulated users remains a central bottleneck constraining the credibility of simulator-based evaluation (Naous et al., 2025).

3 Problem Formulation

Interactive user simulation. We formulate multi-turn user simulation as conditional sequential generation through interaction with a conversational assistant. Let u denote the user profile and task background available before simulation, and let h_t denote the interaction context at turn t, including the dialogue history and the assistant's latest utterance. The simulator policy generates

$$(\hat{\ell}_t, \hat{y}_t) \sim \pi_\theta(\cdot \mid u, h_t), \qquad \hat{\ell}_t = (\hat{r}_t, \hat{z}_t),$$

where $\hat{r}_t$ is a natural-language rationale, $\hat{z}_t$ is an ordinal intent score, and $\hat{y}_t$ is the user's response. The assistant follows a policy π_A and responds to the evolving dialogue, producing the context for the next user turn. A simulated session ends upon conversion, dropout, or reaching the maximum number of turns.

Behavioral trajectory representation. For each observed user response y_t, we associate an annotation $\ell_t = (r_t, z_t)$ consisting of a rationale and an intent score inferred from the dialogue. These annotations provide an interpretable proxy for the user's expressed engagement and commitment. The intent score can increase, decrease, or remain unchanged as the interaction unfolds. We represent an annotated human session and a simulated session as

$$\tau = \{(z_t, y_t)\}_{t=1}^{T}, \qquad \hat{\tau} = \{(\hat{z}_t, \hat{y}_t)\}_{t=1}^{T^*},$$

where T and T^* are their respective lengths, and z_T and $\hat{z}_{T^*}$ denote their final intent scores. Terminal outcomes characterize the observable result of an interaction, such as whether the user provides contact information. This representation supports analysis of intent evolution, response content, and interaction length alongside terminal behavior.

Behavioral modeling objective. Let $p_H(\tau \mid u, h_1, \pi_A)$ denote the conditional distribution of annotated human trajectories, and let $p_\theta(\hat{\tau} \mid u, h_1, \pi_A)$ denote the trajectory distribution induced by the simulator interacting with the assistant. Our goal is to learn π_θ that captures human behavioral dynamics under these conditions, including the distribution of terminal outcomes, the evolution of expressed intent, and the linguistic and temporal characteristics of the interaction. Observed human sessions provide empirical reference trajectories for learning; Section 4 describes the resulting supervision and policy optimization. For evaluation, we aggregate human and simulated trajectories within groups of users with comparable pre-interaction contexts, with each rollout initialized from its own user's context. The grouping procedure and behavioral metrics are specified in Section 5.1.

4 TRACER

Faithful multi-turn user simulation requires both natural responses and coherent behavior across an interaction. Relevant behavioral patterns include the final intent state, the sequence of intermediate intent transitions, and the pace at which the conversation progresses. Fitting individual responses provides no explicit trajectory-level constraint, while terminal supervision alone leaves intermediate behavior underdetermined. This motivates two design questions: how to evaluate behavioral consistency across a session, and how to use that feedback to guide updates at individual turns.

TRACER addresses these questions through two-stage training (Figure 1). Supervised fine-tuning initializes user-like language generation and structured intent prediction. Interactive reinforcement learning then combines outcome and trajectory rewards with deviation-aware advantage modulation. A shared trajectory alignment connects evaluation and optimization: its overall cost scores a completed interaction, while its local discrepancies inform turn-level updates.

4.1 Simulator Initialization and Interactive Rollout

Following Section 3, the simulator generates a rationale r_t, intent score z_t, and user reply y_t conditioned on user information u and context h_t. We initialize these outputs by supervised fine-tuning on real dialogues and their derived annotations with a standard autoregressive likelihood objective.

To train under contexts induced by its own responses, the simulator then interacts with a fixed assistant policy π_A. Each rollout starts from the corresponding user information and initial assistant utterance, and continues until conversion, dropout, or the turn limit. The generated user reply advances the interaction with the assistant. For each input, we sample G trajectories $\{\hat{\tau}_i\}_{i=1}^{G}$ of lengths T_i^*, scored against the paired human reference $\tau = \{(z_t, y_t)\}_{t=1}^{T}$.

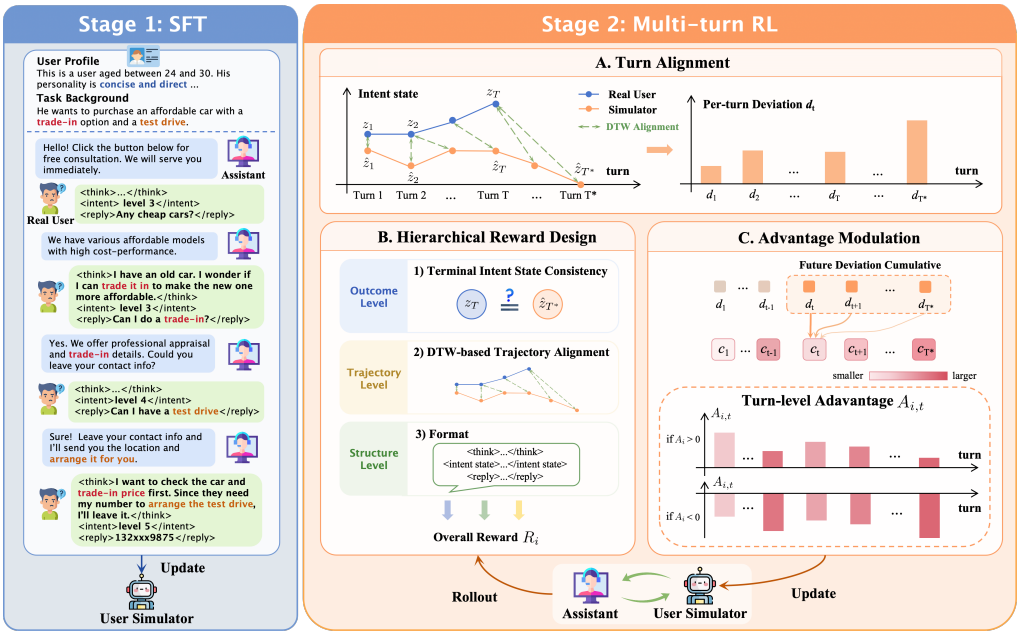


Figure 1: TRACER first learns from annotated human dialogues through supervised fine-tuning, then performs multi-turn reinforcement learning. A shared DTW alignment provides the trajectory cost for session-level rewards and local deviations for turn-level advantage modulation.

4.2 Session-Level Rewards

Behavioral consistency has complementary outcome and process dimensions: a generated session should reflect the reference outcome as well as the intent evolution leading to it. We therefore combine a terminal-intent reward with a trajectory-alignment reward. Together, they provide supervision beyond local response similarity, while a format reward encourages valid structured outputs.

Outcome consistency. The final annotated intent state summarizes the reference session's outcome, providing a direct behavioral target beyond the wording of individual responses. We measure agreement with this state through the terminal-intent reward:

$$R_{\text{final},i} = \mathbb{I}\big(z_T = \hat{z}_{i,T_i^*}\big)\,. \tag{1}$$

This provides an empirical terminal-consistency signal. Terminal agreement alone does not constrain how the interaction unfolds: sessions with the same final intent may exhibit substantially different intent trajectories.

Trajectory alignment. A direct turn-by-turn comparison can penalize comparable intent transitions simply because they occur at different conversational paces. We therefore use dynamic time warping (DTW) (Sakoe & Chiba, 1978; Myers et al., 1980) to align the human and simulated intent sequences while preserving their temporal order. With local cost $|\hat{z}_{i,t} - z_s|$, let $\mathcal{P}_i$ denote the set of warping paths satisfying the standard boundary, monotonicity, and step constraints. The optimal path and its cumulative cost are

$$P_i^* \in \arg\min_{P\in\mathcal{P}_i} \sum_{(t,s)\in P} |\hat{z}_{i,t} - z_s|, \qquad D_i = \sum_{(t,s)\in P_i^*} |\hat{z}_{i,t} - z_s|. \tag{2}$$

Flexible alignment accommodates differences in pacing, but does not by itself require similar session lengths. We therefore combine the alignment cost with an explicit length penalty and a coefficient that discourages nearly constant intent sequences:

$$R_{\text{traj},i} = \exp\left[-\alpha\,\frac{D_i + \beta|T_i^* - T|}{T}\right]\eta_i. \tag{3}$$

Here, $\alpha > 0$ controls reward sensitivity, $\beta \geq 0$ weights the length mismatch, and normalization by T scales the cost relative to the reference length. The coefficient $\eta_i \in (0, 1]$ discounts degenerate, nearly constant intent trajectories. Appendix I gives the alignment details.

Outcome and trajectory agreement do not ensure that individual outputs follow the required structure. We therefore include a format reward to check the prescribed output schema and basic field consistency. If $r_{i,t}^{\text{fmt}} \in \{0, 1\}$ is the turn-level format score, the combined session reward is

$$R_{\text{fmt},i} = \frac{1}{T_i^*} \sum_{t=1}^{T_i^*} r_{i,t}^{\text{fmt}}, \quad R_i = \lambda_1 R_{\text{fmt},i} + \lambda_2 R_{\text{final},i} + \lambda_3 R_{\text{traj},i}. \tag{4}$$

The resulting scalar incorporates output validity, terminal agreement, and intermediate trajectory information. Reward hyperparameters are reported in Appendix A.

4.3 Deviation-Aware Policy Optimization

Following GRPO (Shao et al., 2024), we standardize rewards within each rollout group to obtain the trajectory advantage A_i. Broadcasting this scalar across turns does not reflect their different alignment deviations. We therefore retain its global reinforcement direction while adapting update strength using the shared DTW alignment.

Cumulative alignment deviation. We reuse the optimal DTW path P_i^* from Eq. equation 2 to derive a local deviation for each generated turn. With $\mathcal{A}_{i,t} = \{s : (t, s) \in P_i^*\}$ denoting the reference turns aligned to generated turn t, we define

$$d_{i,t} = \frac{1}{|\mathcal{A}_{i,t}|} \sum_{s \in \mathcal{A}_{i,t}} |\hat{z}_{i,t} - z_s|. \tag{5}$$

This averages the discrepancies over all reference turns aligned to the same generated turn. A local score describes agreement at one position, but does not indicate whether the subsequent interaction remains aligned. To incorporate this continuation information, we aggregate deviations over a discounted suffix of the sampled trajectory:

$$c_{i,t} = \sum_{k=0}^{T_i^* - t} \rho^k d_{i,t+k}, \qquad \rho \in [0, 1]. \tag{6}$$

The discount ρ controls how strongly later discrepancies affect the score: smaller values emphasize nearby turns, while larger values retain more information about the continuation. Setting $\rho = 0$ recovers the local signal $d_{i,t}$.

Advantage modulation. A relatively high-reward session can still contain poorly aligned segments, while a low-reward session may exhibit varying degrees of deviation across turns. We use $c_{i,t}$ to adjust the strength of reinforcement or suppression within each session, with a bounded modulation that preserves the sign of the trajectory advantage:

$$\widetilde{A}_{i,t} = w_{i,t} A_i, \qquad w_{i,t} = 1 - \lambda_{\text{adv}} \operatorname{sign}(A_i) \tanh(\gamma c_{i,t}), \tag{7}$$

where $\lambda_{\text{adv}} \in [0, 1]$ bounds the modulation magnitude and $\gamma > 0$ controls sensitivity. Equivalently, $\widetilde{A}_{i,t} = A_i - \lambda_{\text{adv}} |A_i| \tanh(\gamma c_{i,t})$. Larger cumulative deviations reduce reinforcement when $A_i > 0$ and increase the magnitude of suppression when $A_i < 0$. The modulation prevents sign reversal, while $\lambda_{\text{adv}} = 0$ recovers the unmodulated trajectory advantage.

We use $\widetilde{A}_{i,t}$ in place of A_i for all simulator tokens selected for optimization at turn t, excluding assistant and padding tokens. Alignment and advantage values remain fixed during each update. This preserves the GRPO clipped surrogate structure while changing the turn-level optimization signal.

5 Experiments

5.1 Experimental Setup

Datasets. We construct CustomerService-Dialogue from real-world dialogues between users and merchant-side customer service agents on a mobile application, spanning industries such as health-

care and automotive. Each session contains a 3–15-turn dialogue, a user profile (demographics, personality, and initial intent), and merchant information. Conversion is defined by whether the user provides contact information. Claude-4.5-Sonnet provides weak turn-level supervision $\ell_t = (r_t, z_t)$, generating a rationale and an intent score conditioned on the full dialogue and user profile. Human assessment on a stratified sample shows strong annotation agreement (Table 4). The dataset contains 8,040 training sessions and 3,866 test sessions organized into 762 reference groups. Each group contains sessions from the same merchant, with similar personality traits and initial user needs selected using semantic similarity thresholds. Grouping is used only for evaluation; training retains paired-session supervision. Construction and grouping details appear in Appendix B.

Evaluation Protocol. For each test instance, the simulated user model receives the user profile and the assistant's first utterance. It role-plays as the user and conducts a multi-turn interaction with the assistant until the dialogue terminates via conversion, dropout or reaching a maximum turns. The assistant is the same model deployed online, conditioned on the merchant information and user query, and follows a standardized SOP-driven policy that yields low-variance, reproducible responses. Metrics are computed over the simulated trajectories against the ground-truth dialogues.

Metrics. We evaluate three dimensions. **Session-Level Behavior**: PCR reports the overall conversion rate; ACC and conversion-class F1 measure outcome agreement with the original paired sessions. **Group-Level Fidelity**: Group-ΔPCR measures the absolute difference between simulated and real conversion rates within each group; W1-Turns measures the Wasserstein-1 distance between their dialogue-length distributions. **Semantic Content**: OT-DTW measures optimal transport distance between simulated and real sessions within each group, using DTW-based semantic trajectory costs. Group-based metrics are averaged across the 762 groups.

Baselines. We compare three categories of user simulators. **General LLM baselines:** We evaluate several general LLMs as zero-shot user simulators, including the Qwen2.5-7B-Instruct(Qwen et al., 2025), Qwen3-4B-Instruct(Yang et al., 2025), DeepSeek-V3.2(Liu et al., 2025), Gemini-3.1-Pro, Doubao-Seed-2.0-Pro and Claude-4.5-Sonnet. **Role-playing baselines:** We include models explicitly designed for role-playing, including Doubao-Seed-Character, Qwen-Character, and CharGLM-4, which emphasize strong role consistency and conversational naturalness. **Training-based baseline:** We adopt HUMANLM (Wu et al., 2026), which introduces latent natural-language states aligned with responses via RL. We train it on our dataset using SFT and RL for fair comparison.

Implementation details. We instantiate TRACER using Qwen2.5-7B-Instruct and Qwen3-4B-Instruct. In the Stage 1, we train the models for 2 epochs with a learning rate of 10^{-5}. In Stage 2, we train for 6 epochs, sample 8 rollout trajectories per prompt. All models use an inference temperature of 0.7. Full configurations, evaluation details, and stability analysis are provided in Appendix A.

5.2 Main Results

Table 1 reports the performance of all methods along the three evaluation dimensions. We summarize our main findings as follows.

General LLMs exhibit insufficient behavioral alignment, with opposing tendencies toward over-compliance and over-refusal. Their PCR ranges from 5.8% to 79.7% , compared with the real conversion rate of 51.24%, while ACC remains within 50–59% and F1 reaches at most 68.2%. These results suggest that general LLMs tend to be either overly willing or overly reluctant to convert, failing to match the conversion behavior observed among real users.

Role-playing baselines remain behaviorally misaligned despite their focus on persona consistency. Doubao-Seed-Character, Qwen-Character, and CharGLM-4 exhibit a pronounced under-conversion tendency, with PCRs of 9.9–26.2% and F1 scores of 23.9–37.0%. Their high W1-Turns and OT-DTW further reveal mismatches in dialogue length and semantic progression. These results suggest that persona-oriented modeling alone is insufficient to reproduce real users' conversion tendencies and multi-turn interaction patterns.

TRACER jointly achieves linguistic-style consistency and trajectory-level behavioral consistency. TRACER-7B achieves a PCR of 50.5% and outperforms all baselines on ACC, F1, Group-ΔPCR, and OT-DTW, exceeding the strongest baseline by 11.4 F1 points. Its W1-Turns performance ranks second,

Table 1: Overall comparison of user simulators across three evaluation dimensions: Session-Level Behavior (PCR, ACC, F1), Group-Level Fidelity (Group-ΔPCR, W1-Turns), and Semantic Content (OT-DTW). Groups are balanced 1:1; the session-level real PCR is 51.24%. Best results are in bold, second-best are underlined.

Method	Session-Level Behavior			Group-Level Fidelity		Semantic Content
	PCR	ACC ↑	F1 ↑	Group-ΔPCR ↓	W1-Turns ↓	OT-DTW ↓
General Baselines						
Qwen2.5-7B-Instruct	36.7	53.7	47.4	0.456	5.383	0.500
Qwen3-4B-Instruct	32.6	57.8	49.6	0.419	5.307	0.505
Gemini-3.1-Pro	40.8	56.4	52.6	0.425	1.588	0.480
Doubao-Seed-2.0-Pro	79.7	58.4	68.2	0.425	2.633	0.466
Claude-4.5-Sonnet	5.8	50.7	13.7	0.473	2.278	0.480
Deepseek-V3.2	33.7	55.3	47.4	0.426	2.689	0.486
Role-playing Baselines						
Doubao-Seed-Character	26.2	51.3	37.0	0.471	7.063	0.531
Qwen-Character	9.9	53.5	23.9	0.453	8.618	0.521
CharGLM-4	16.6	51.5	28.5	0.470	7.549	0.529
Trained on Our Dataset						
HUMANLM-8B	13.2	57.4	33.9	0.390	**0.944**	0.443
TRACER-7B	50.5	<u>79.2</u>	<u>79.6</u>	<u>0.165</u>	<u>0.977</u>	**0.358**
TRACER-4B	51.6	**79.8**	**80.3**	**0.161**	1.43	<u>0.366</u>

behind HUMANLM. Among all evaluated models, the 4B variant achieves the best performance on ACC, F1, and Group-ΔPCR, while the 7B variant performs best on OT-DTW. These results support the framework's effectiveness across both backbones. Compared with HUMANLM trained on the same data using SFT and RL, both variants achieve better outcome agreement and semantic trajectory alignment, highlighting the value of jointly aligning terminal outcomes and the course of interaction.

5.3 Detailed Analysis

5.3.1 Human Evaluations

To assess simulation fidelity, we conduct a Turing-style evaluation testing whether AI-simulated users are distinguishable from real humans. Annotators are given two dialogues with the same user profile information, one produced by a real user and the other by a simulator, and are asked to identify the human-generated dialogue. We randomly sample 500 sessions from the test set, covering diverse personas and intents. Details about the annotators are provided in the Appendix H.1.

TRACER-7B achieves human-level indistinguishability. Figure 2 compares TRACER-7B with selected general-purpose and role-playing baselines from Table 1, as well as its SFT-only variant. General LLMs are correctly identified in over 98% of cases, revealing persistent machine-like patterns. The SFT variant reduces identification accuracy to 61.80% but still falls short of indistinguishability. In contrast, TRACER-7B reaches 49.50%, converging to the random-guess threshold and indicating that it captures subtle human behavioral nuances missed by prior simulators.

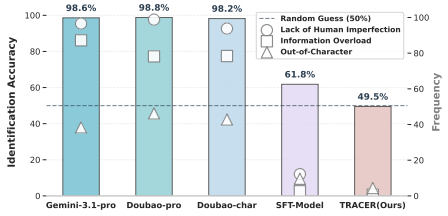


Figure 2: Human-indistinguishability evaluation. Accuracy is shown as bars, while the presence of non-human characteristics is indicated with markers. The dashed line at 50% represents the random guess threshold.

Why are baselines recognized? Table 6 groups detection cues into three dimensions. (1) *Linguistic over-regularity*: baselines almost entirely lack human imperfection and frequently show overly neat structure. (2) *Mis-calibrated interaction rhythm*: they are consistently information-overloading and excessively cooperative, with Doubao variants further exhibiting mechanical patterns. (3) *Out-of-character* behavior appears in roughly 40% of cases, while higher-order defects such as emotional discontinuity remain rare. Details are in Appendix E.

5.3.2 Ablation Study

Table 2: Ablation study of TRACER-7B on session-level behavior, dialogue structure, and semantic content. Results are reported as mean ± standard deviation over five evaluation seeds.

Method	Session-Level Behavior			Group-Level Fidelity		Semantic Content
	PCR	ACC ↑	F1 ↑	Group-ΔPCR ↓	W1-Turns ↓	OT-DTW ↓
TRACER-7B	50.3 ± 0.2	$\mathbf{79.0} \pm 0.2$	$\mathbf{79.4} \pm 0.2$	$\mathbf{0.166} \pm 0.003$	$\mathbf{0.972} \pm 0.006$	$\mathbf{0.357} \pm 0.001$
– w/o advantage modulation	49.5 ± 0.4	$\underline{78.7} \pm 0.1$	$\underline{78.8} \pm 0.2$	$\underline{0.181} \pm 0.004$	$\underline{1.030} \pm 0.014$	0.369 ± 0.002
– w/o R_{traj}	52.8 ± 0.3	77.4 ± 0.3	78.3 ± 0.3	0.187 ± 0.005	1.036 ± 0.020	0.373 ± 0.001
– w/o DTW alignment	48.3 ± 0.3	77.7 ± 0.3	77.6 ± 0.4	0.186 ± 0.004	1.064 ± 0.014	0.367 ± 0.002
– w/o multi-turn RL	52.8 ± 0.4	74.3 ± 0.3	75.3 ± 0.2	0.218 ± 0.004	1.081 ± 0.007	$\underline{0.365} \pm 0.001$
– w/o RL	12.2 ± 0.5	57.6 ± 0.1	33.2 ± 0.4	0.398 ± 0.003	1.962 ± 0.035	0.380 ± 0.001

Table 2 summarizes the ablation results for TRACER. The full model achieves the best mean performance across all five direction-specified metrics. Removing advantage modulation or the trajectory-level reward R_{traj} consistently degrades behavioral accuracy, group-level fidelity, and semantic alignment, supporting the contribution of each component. Replacing DTW with strict turn-by-turn matching also worsens all five metrics, supporting the value of accommodating differences in conversational pacing. Disabling multi-turn RL substantially reduces behavioral accuracy and worsens Group-ΔPCR, highlighting the value of multi-turn optimization. Removing RL entirely causes the largest overall degradation, underscoring its importance beyond supervised fine-tuning.

5.3.3 Out-of-Domain Generalization

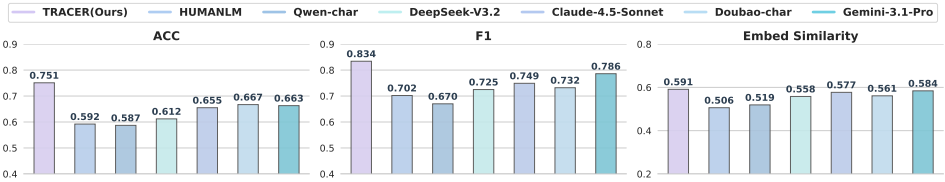


Figure 3: Performance Comparison of Models on the Out-of-Domain Dataset.

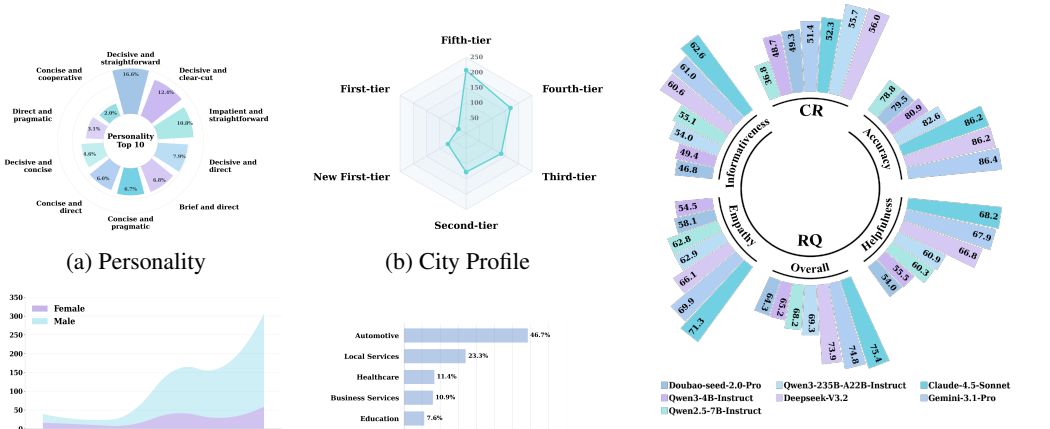
To assess TRACER's generalization beyond its training distribution, we evaluate it on CSC-Conv (Zhu et al., 2026), an open-source dataset of real user–agent conversations in financial customer service unseen during training. For each dialogue, we extract the user's initial intent and a binary resolution label defining the session-level terminal state. We randomly sample 1,000 dialogues and compare TRACER with the strongest baselines from Table 1 using terminal-state ACC, F1, and dialogue embedding similarity. As shown in Figure 3, TRACER performs best on all three metrics, outperforming the strongest baseline by +8.4, +4.8, and +0.7 points, respectively, despite no exposure to this domain. These outcome- and trajectory-level gains suggest TRACER learns transferable user-simulation behaviors rather than domain-specific patterns. Detailed settings are provided in Appendix F.

6 Dynamic Marketing Benchmark

TRACER aligns simulated users with human trajectories, capturing intent changes and terminal decisions. This enables multi-turn assistant evaluation, where user responses shape interactions and decisions provide outcome feedback. We instantiate this as the Dynamic Marketing Benchmark, comparing response quality and simulated conversion under shared user and merchant conditions.

6.1 Benchmark Construction

DM-Bench comprises 734 instances sampled from CustomerService-Dialogue. Each instance contains a user profile, including demographic attributes and initial intent, together with merchant information. Seven assistant models interact with TRACER under the same set of initial conditions, with user responses and intent evolving as each conversation unfolds. Figure 4 summarizes the user and merchant distributions. The data source and construction protocol are described in Section 5.1.



(a) Personality (b) City Profile (c) Demographic Profile (d) Merchant Distribution (e) LLM Performance as Customer Service Agents

Figure 4: Overview of DM-Bench: User Profiles, Merchant Distribution, and Model Performance.

We report two complementary metrics. **Simulated Conversion Rate (CR)** measures the proportion of sessions ending in conversion, as determined by the user simulator. **Response Quality (RQ)** follows the LLM-as-a-judge protocol of Zhu et al. (2026), assessing multi-turn responses in terms of accuracy, helpfulness, informativeness, and empathy. Together, these metrics capture response quality and terminal outcomes within the simulated environment.

Behavioral validation. We validate feedback through turn-level agreement with human expectations and assistant-ranking agreement with human role-play. For intent changes, three annotators label 449 instances from 200 conversations across six strategy types (pairwise agreement $> 90\%$). Against majority-vote labels, TRACER achieves recalls of 88.7%, 74.2%, and 81.0% for decrease, unchanged, and increase, respectively (Appendix G.2). Human role-play on 100 contexts covers seven assistants (Table 9). Human and TRACER conversion-rate rankings show strong agreement (Spearman's $\rho = 0.955$; Kendall's $\tau_b = 0.878$). These results support the plausibility of TRACER's local intent changes and its agreement with human role-play in ranking assistants.

6.2 Results and Findings

We report the results in Figure 4, Table 8 and Table 10, with the following conclusions: **(1) Response quality is not a reliable proxy for conversion effectiveness.** RQ and CR show only moderate rank correlation across seven models (Spearman $\rho = 0.57$, $p = 0.18$, $n = 7$). For instance, Claude-4.5-Sonnet ranks highest on RQ but is surpassed in CR by DeepSeek-V3.2 and Qwen3-235B, indicating that response quality does not fully capture persuasion effectiveness. **(2) Even frontier LLMs face a substantial conversion ceiling.** Despite high RQ scores, top models still fail to convert many users, suggesting limited adaptation to user intent and decision dynamics across turns.

These findings show the value of DM-Bench as a user-centered, outcome-driven evaluation protocol: it reveals performance gaps that are largely invisible under quality-only benchmarks, especially the gap between generating good-sounding responses and generating decision-moving ones.

7 Conclusion

In this work, we present TRACER, a multi-turn user simulator aligned with real users at both linguistic and decision-making levels, trained via a two-stage paradigm of SFT and trajectory-aligned RL. Building on TRACER, we further construct the Dynamic Marketing Benchmark, which jointly measures persuasion effectiveness and response quality. Experiments show that TRACER largely outperforms existing simulators and achieves near-human fidelity in Turing tests, while revealing a misalignment between response quality and persuasion effectiveness in mainstream LLMs. In future work, we plan to extend TRACER to broader interactive scenarios and explore its use as a training environment for dialogue agents.

AI USE STATEMENT

During the preparation of this study, LLMs were utilized solely as auxiliary tools. The application of these models was strictly limited to linguistic refinement, including grammar and spelling corrections, as well as technical support for drafting and debugging code during the experimental phase. All suggestions provided by the models underwent rigorous auditing, revision, and verification by the authors. It must be explicitly stated that LLMs played no role in the conceptualization of the research problem, the development of the theoretical framework, the design of experimental protocols, or the scientific interpretation of the results. All core academic contributions such as the research logic, methodological design, and final conclusions were independently completed by the whole authors.

ETHICS STATEMENT

All user behavior data involved in this study were sampled and processed in strict accordance with relevant laws, regulations, platform guidelines, and privacy protection requirements. The scope of this research is limited exclusively to data resources authorized for scientific analysis and benchmark construction. Prior to the release of the public benchmark, the raw data underwent multiple stages of quality control, representative sampling, and de-identification procedures. Sensitive information possessing identifying characteristics, such as names, contact details, and specific addresses, has been either deleted or replaced with semantic placeholders. Furthermore, this study implemented filters for inappropriate or potentially harmful content and conducted manual audits to ensure that the processed data meet ethical standards. All analyses and experiments were performed using the desensitized and cleaned dataset. The authors assume full legal and academic responsibility for ensuring that the construction and utilization of the benchmark data comply with all relevant data governance and privacy protection norms.

REPRODUCIBILITY STATEMENT

To support reproducibility of our experiments, we provide the TRACER code at `https://anonymous.4open.science/r/TRACER_code-0A27` and document the experimental procedures in the paper and appendices. Section 4 specifies the training objectives and optimization procedure, while Appendix A reports the hardware, training and inference configurations, and hyperparameter settings. Appendix B describes data construction and preprocessing, and Appendix K provides the prompts used for annotation and simulation. The evaluation protocol is presented in Section 5.1, with additional details on out-of-domain evaluation, DM-Bench, and human annotation in Appendices F– H.

REFERENCES

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 8–17, 2024.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35212–35290, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1786. URL `https://aclanthology.org/2025.emnlp-main.1786/`.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3953–3957, 2023.

Jaehyung Kim and Yiming Yang. Few-shot personalization of llms with mis-aligned responses. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11943–11974, 2025.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30084–30099, 2025.

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

Chonghua Liao, Ke Wang, Yuchuan Wu, Fei Huang, and Yongbin Li. Moa: Multi-objective alignment for role-playing agents. *arXiv preprint arXiv:2512.09756*, 2025.

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.

Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Qi He, Dakuo Wang, et al. Can llm agents simulate multi-turn human behavior? evidence from real online customer behavior data. *arXiv preprint arXiv:2503.20749*, 2025.

C. Myers, L. Rabiner, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28 (6):623–635, 1980. doi: 10.1109/TASSP.1980.1163491.

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

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Zhiwei Liu, Jianguo Zhang, Haolin Chen, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, et al. Userbench: An interactive gym environment for user-centric agents. *arXiv preprint arXiv:2507.22034*, 2025a.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Jielin Qiu, Zhiwei Liu, Haolin Chen, Shirley Kokane, Heng Ji, Weiran Yao, Shelby Heinecke, et al. Userrl: Training interactive user-centric agent via reinforcement learning. *arXiv preprint arXiv:2509.19736*, 2025b.

Huachuan Qiu and Zhenzhong Lan. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*, 2024.

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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008.

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.

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.

Yimin Shi, Yang Fei, Shiqi Zhang, Haixun Wang, and Xiaokui Xiao. You are what you bought: Generating customer personas for e-commerce applications. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1810–1819, 2025.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 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)*, pp. 21082–21107, 2025a.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, et al. Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *arXiv preprint arXiv:2507.03112*, 2025b.

Zongsheng Wang, Kaili Sun, Bowen Wu, Qun Yu, Ying Li, and Baoxun Wang. Raiden-r1: Improving role-awareness of llms via grpo with verifiable reward. *arXiv preprint arXiv:2505.10218*, 2025c.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. In *International Conference on Machine Learning*, pp. 67260–67283. PMLR, 2025.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. Humanlm: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026.

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.

Haojin Yang, Ai Jian, Xinyue Huang, Yiwei Wang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Jingqing Ruan. Harmonizing dense and sparse signals in multi-turn rl: Dual-horizon credit assignment for industrial sales agents. *arXiv preprint arXiv:2603.01481*, 2026.

Xinge Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization, 2025. URL https://arxiv.org/abs/2508.09074.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, et al. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*, 2025.

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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

Jie Zhu, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, Chi Zhang, and Fang Kong. Evaluating, synthesizing, and enhancing for customer support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35185–35194, 2026.

APPENDIX

A EXPERIMENTAL DETAILS

A.1 IMPLEMENTATION DETAILS

Training configuration. All training experiments were conducted on a hardware platform equipped with $8\times$ NVIDIA H200 GPUs. For the backbone models, we mainly adopted Qwen2.5-7B-Instruct (Qwen et al., 2025) and Qwen3-4B-Instruct (Yang et al., 2025) as the base models, and trained them through supervised fine-tuning and multi-turn reinforcement learning, respectively.

During the supervised fine-tuning stage, we implemented the SFT training pipeline based on LLaMA-Factory (Zheng et al., 2024) and used DeepSpeed ZeRO-2 for distributed training. The maximum context length was set to 4096 tokens. The per-GPU batch size was set to 4, with a gradient accumulation step of 2. The models were trained for 2 epochs with a learning rate of 1.0×10^{-5}. We adopted a cosine learning rate scheduler with a warmup ratio of 0.1. All SFT experiments were conducted using bf16 precision to improve training efficiency and reduce GPU memory consumption.

During the multi-turn reinforcement learning stage, we modified the verl framework (Sheng et al., 2024) to construct a multi-turn RL interaction training framework tailored to the user simulator dialogue learning task. The global training batch size was set to 32, the mini-batch size was set to 32, and the per-GPU micro-batch size was set to 4. For each prompt, we sampled 8 rollout trajectories. The total number of training epochs was set to 6, and the sampling temperature was set to 1.0. On the input side, the maximum prompt length was set to 6144 tokens, while the maximum response length was set to 2048 tokens. We further limited the maximum number of dialogue turns during rollout to 15. Under the above configuration, multi-turn RL training took approximately 12 hours.

Inference configuration. During the inference stage, the temperature was set to 0.7 for both the baseline models and our model to ensure a consistent decoding configuration across all compared methods. Following the dataset setting, the maximum number of interaction turns was capped at 15 during inference, and sessions that did not reach conversion within this limit were treated as dropout.

Evaluation details. Across all evaluations, we employ a unified computational setup to ensure comparability among different methods. All token-related metrics are computed using the Qwen2.5-7B-Instruct tokenizer.

Group-ΔPCR is the mean absolute difference between simulated and real conversion rates across reference groups: $\text{Group-}\Delta\text{PCR} = G^{-1} \sum_{g=1}^{G} |\hat{p}_g - p_g|$, where $\hat{p}_g$ and p_g denote the simulated and real session-level conversion rates in group g. Groups receive equal weight, and the metric is reported as a proportion rather than a percentage.

W1-Turns averages the Wasserstein-1 distance between the empirical distributions of real and generated user-turn counts within each group, giving all groups equal weight. It is measured in user turns. Real turns are counted from non-empty user messages after merging consecutive messages from the same speaker; generated turns are counted from the interaction log, including valid empty exit replies.

For OT-DTW, non-empty user replies are encoded using bge-base-zh-v1.5 and L2-normalized, excluding reasoning text and assistant replies. The local cost is $\max(0, 1 - \cos(\mathbf{e}_i, \mathbf{e}_j))$. For each generated–real trajectory pair, we find the minimum cumulative-cost DTW path and divide its cost by its length, rather than directly minimizing average cost. The cost is 1 if exactly one trajectory is empty and 0 if both are empty. Within each group, the Hungarian algorithm finds a minimum-cost one-to-one matching over the pairwise trajectory costs. Matched costs are averaged within groups, followed by an equally weighted average across groups. Lower values indicate better agreement for all three metrics.

For Embedding Similarity in the out-of-domain evaluation, we utilize bge-base-zh-v1.5 as the encoding model to map both simulated and ground-truth user replies into normalized embeddings, measuring their semantic consistency via cosine similarity.

Table 3: Hyperparameter sensitivity of TRACER-7B. Each configuration is evaluated using its corresponding fixed checkpoint with five evaluation seeds (42–46); hyperparameter variants correspond to separately trained checkpoints. Results are mean ± standard deviation across evaluation seeds, rather than independent training runs. PCR, ACC, and F1 are reported in percent. The reference PCR is 51.24%. Bold indicates the best mean for each directional metric.

Configuration	λ_{adv}	γ	ρ	PCR	ACC ↑	F1 ↑	Group-ΔPCR ↓	W1-Turns ↓	OT-DTW ↓
Without advantage modulation	–	–	–	49.54 ± 0.44	78.68 ± 0.13	78.85 ± 0.19	0.181 ± 0.004	1.030 ± 0.014	0.369 ± 0.002
Default TRACER-7B	0.2	0.5	0.4	50.29 ± 0.20	79.04 ± 0.19	79.36 ± 0.21	0.166 ± 0.003	$\mathbf{0.972} \pm 0.006$	$\mathbf{0.357} \pm 0.001$
$\rho = 0.0$	0.2	0.5	0.0	50.52 ± 0.32	78.36 ± 0.20	78.74 ± 0.15	0.176 ± 0.002	0.994 ± 0.012	0.373 ± 0.001
$\rho = 0.1$	0.2	0.5	0.1	47.21 ± 0.22	79.06 ± 0.24	78.73 ± 0.26	0.164 ± 0.003	0.975 ± 0.007	0.369 ± 0.001
$\rho = 0.3$	0.2	0.5	0.3	48.66 ± 0.19	78.55 ± 0.29	78.53 ± 0.28	0.170 ± 0.004	0.977 ± 0.017	0.365 ± 0.001
$\rho = 0.5$	0.2	0.5	0.5	53.86 ± 0.16	78.42 ± 0.24	79.46 ± 0.23	0.177 ± 0.003	0.987 ± 0.009	0.365 ± 0.002
$\lambda_{\text{adv}} = 0.1$	0.1	0.5	0.4	47.78 ± 0.32	$\mathbf{79.80} \pm 0.20$	$\mathbf{79.60} \pm 0.25$	$\mathbf{0.160} \pm 0.002$	1.059 ± 0.012	0.382 ± 0.002
$\lambda_{\text{adv}} = 0.3$	0.3	0.5	0.4	52.16 ± 0.25	78.47 ± 0.29	79.18 ± 0.25	0.172 ± 0.005	1.101 ± 0.024	0.371 ± 0.001
$\gamma = 0.4$	0.2	0.4	0.4	50.82 ± 0.18	78.67 ± 0.11	79.10 ± 0.07	0.168 ± 0.002	0.989 ± 0.012	0.361 ± 0.002

Reward and advantage hyperparameters. Unless otherwise specified, we use the same reward and advantage modulation configurations across both model scales. All trajectory-level rewards are computed at the session level and propagated to valid user turns via the proposed advantage modulation strategy. For the DTW-based trajectory reward in Eq. equation 3, we set the path-deviation sensitivity coefficient to $\alpha = 1.0$ and the length-penalty coefficient to $\beta = 1.0$. The flattening coefficient is set to $\eta = 0.2$ when the predicted intent trajectory degenerates into an overly flat sequence, and $\eta = 1.0$ otherwise. For the hierarchical reward in Eq. equation 4, we set $(\lambda_1, \lambda_2, \lambda_3) = (0.2, 0.5, 0.3)$ for the format reward, terminal intent consistency reward, and trajectory alignment reward, respectively. For the cumulative future deviation score in Eq. equation 6, we set the future-deviation discount factor to $\rho = 0.4$. For the turn-level advantage modulation in Eq. equation 7, we set the modulation strength to $\lambda_{\text{adv}} = 0.2$, the deviation sensitivity coefficient to $\gamma = 0.5$, and the numerical stability term to $\epsilon = 10^{-6}$. All coefficients were selected on a held-out validation split.

A.2 Sensitivity Analysis of Hyperparameters in Multi-Turn Advantage Modulation

Overall sensitivity. Table 3 varies one hyperparameter at a time around the default setting $(\lambda_{\text{adv}}, \gamma, \rho) = (0.2, 0.5, 0.4)$. All tested configurations with advantage modulation achieve lower mean Group-ΔPCR than the unmodulated baseline, but improvements in ACC and F1 are not uniform. Their ACC and F1 remain within 78.36–79.80 and 78.53–79.60, respectively, while PCR varies more substantially, from 47.21% to 53.86%. The reported standard deviations indicate limited variation across evaluation seeds for each fixed checkpoint; they do not measure variability across independent training runs.

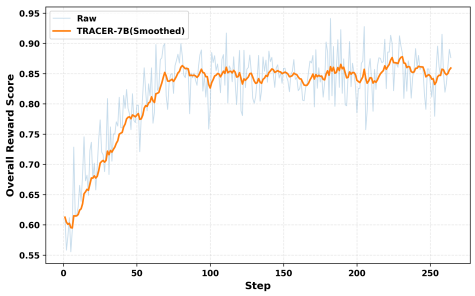
Future-deviation discount ρ. Setting $\rho = 0$ restricts modulation to the current turn's local deviation. Compared with this setting, the default $\rho = 0.4$ improves mean performance on all five directional metrics. The effects are not monotonic: $\rho = 0.1$ obtains slightly better ACC and Group-ΔPCR than the default, whereas $\rho = 0.5$ achieves higher F1 but worse W1-Turns. Among the tested discount values, the default achieves the best performance on W1-Turns and OT-DTW, indicating better agreement in dialogue-length distributions and semantic trajectories under these metrics.

Modulation strength λ_{adv}. Reducing λ_{adv} to 0.1 yields the highest ACC and F1 and the lowest Group-ΔPCR, but increases both W1-Turns and OT-DTW relative to the default. Increasing λ_{adv} to 0.3 produces lower ACC and F1 and higher values of all three distance metrics than the default. These results reveal a trade-off between outcome agreement and dialogue-length and semantic alignment. The default strength provides a compromise, achieving better trajectory fidelity than the smaller value while retaining improved outcome agreement over the unmodulated baseline.

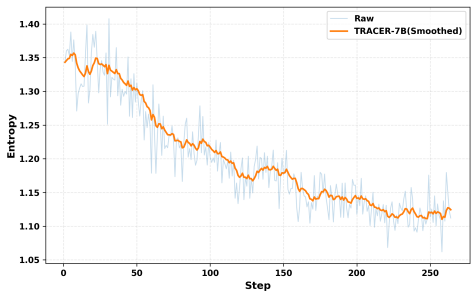
Deviation sensitivity γ. Reducing γ from 0.5 to 0.4 produces similar results, with slightly lower ACC and F1 and slightly higher values of the three distance metrics. This comparison suggests limited sensitivity to this particular perturbation, although the two tested values do not establish

robustness over a broader range of γ. Overall, the default configuration achieves the lowest OT-DTW among all tested configurations while improving all five directional metrics over the unmodulated baseline.

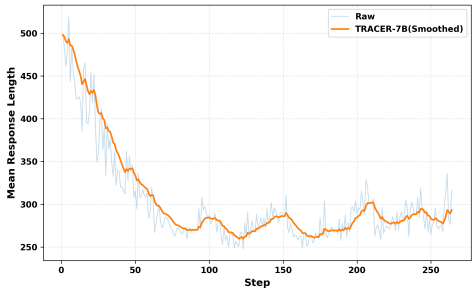
A.3 Training Dynamics and Convergence Analysis



(a) Overall reward score



(b) Generation entropy



(c) Mean response length

Figure 5: Training dynamics of TRACER-7B. We report the overall reward score, generation entropy, and mean response length during reinforcement learning. The solid curves show EMA-smoothed trends, while the faint curves denote the raw values.

Figures 5 and 6 present the training dynamics of TRACER-7B and TRACER-4B, respectively. Across both model scales, the overall reward score increases rapidly during the early stage of reinforcement learning and then gradually stabilizes, indicating that the proposed training objective can be effectively optimized. The convergence pattern is consistent for both 7B and 4B models, suggesting that the proposed framework is not limited to a single model size.

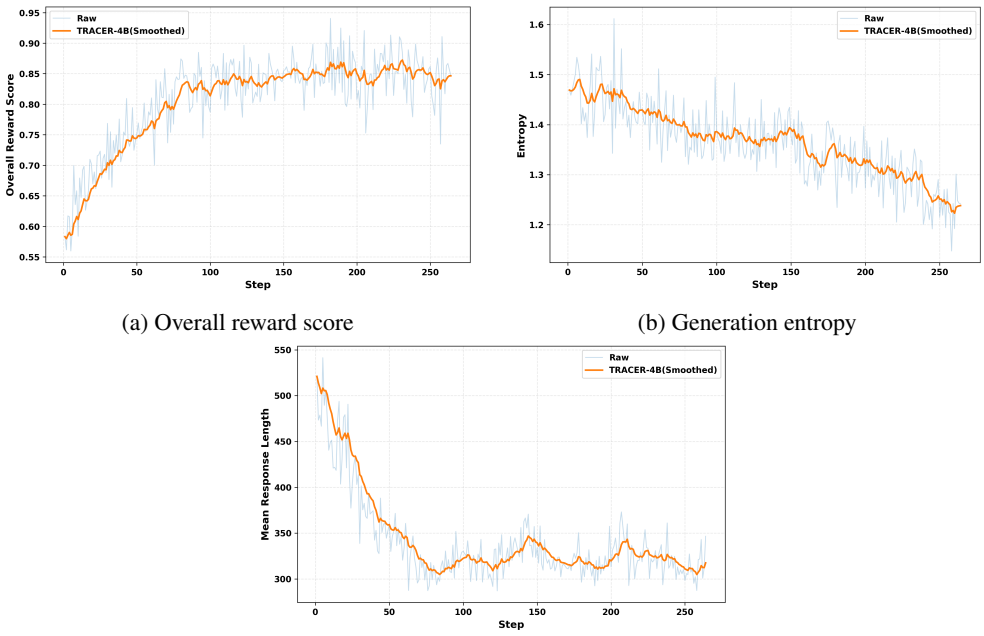
The generation entropy exhibits a smooth downward trend for both models. This pattern suggests that the policy gradually becomes more focused as training progresses, while avoiding abrupt entropy collapse. In other words, the model learns to produce more task-aligned responses without showing signs of severe mode collapse or unstable policy updates.

The mean response length also decreases substantially in the early training stage and then remains within a relatively stable range. This trend is important because it suggests that the reward improvement is not achieved by simply generating increasingly longer responses. Instead, both models learn to produce more concise and controlled outputs after reinforcement learning. Overall, these training curves demonstrate the stability and feasibility of the proposed reinforcement learning framework across model scales, with stable reward optimization, gradual policy adaptation, and controlled response length.

B Dataset Construction Details

B.1 Data Source

The dataset used in this study is derived from real-world merchant–user dialogue logs collected from an online commercial service platform. The raw corpus contains 36,841 multi-turn dialogue



(a) Overall reward score

(b) Generation entropy

(c) Mean response length

Figure 6: Training dynamics of TRACER-4B. We report the overall reward score, generation entropy, and mean response length during reinforcement learning. The solid curves show EMA-smoothed trends, while the faint curves denote the raw values.

sessions between merchant-side intelligent customer service agents and real users, together with corresponding user profile attributes and merchant-side information. The dialogues cover multiple real-world business domains, including healthcare, automotive services, education consulting, and legal consulting.

Before entering the research pipeline, all data were strictly anonymized and de-identified. We removed personally identifiable information and retained only task-relevant structured attributes, merchant-side metadata, and dialogue content required for user simulation. The processed dataset is intended to support user behavior modeling in multi-turn customer service scenarios while minimizing privacy risks.

B.2 User Profile Construction

To support high-fidelity simulation grounded in user background information, we construct structured user profiles from two complementary perspectives: objective attributes and subjective behavioral priors.

First, we organize objective user attributes into natural-language profile descriptions. Based on real business data, we select three groups of core information. The first group consists of basic demographic attributes, including age group, gender, education level, marital or parenting status, and geographic location. The second group contains occupation-related information. The third group reflects consumption-related attributes, including mobile operating system, device price range, city tier, and residential environment. We then convert these discrete labels into coherent natural-language background descriptions using predefined templates. This step provides the user simulator with stable prior conditions while avoiding direct exposure of unnecessary raw attribute fields.

Second, since objective attributes alone are insufficient to determine a user's conversational style, task motivation, and behavioral tendency, we use Claude 4.5-Sonnet as an annotation model to extract subjective behavioral priors. Specifically, with final conversion outcomes and other terminal-state information masked, we analyze early dialogue context and observable user information to extract two types of subjective priors. The first type is the user's personality trait, such as impatient and

direct, rational and cautious, or cooperative and efficient. The second type is the initial core intent, which summarizes the user's primary task goal, key constraints, and expected issue to be resolved at the beginning of the conversation.

To avoid label leakage, we impose strict constraints during subjective information extraction. The extracted information is not allowed to include any signal related to the final conversion outcome, such as whether the user eventually provides contact information. In other words, the initial core intent describes only the user's demand and task background at the beginning of the conversation, without revealing the final session outcome. In this way, the user profile serves as a prior condition for simulation while avoiding direct leakage of the target decision label.

B.3 Dialogue Augmentation and Structured Trajectory Construction

The original real-world dialogue logs mainly contain users' surface-level natural-language responses, but lack explicit annotations of the latent decision states underlying those responses. This leads to a key limitation: if a model is trained only on raw user responses, it primarily learns *what the user says*, but receives limited supervision regarding *why the user responds in that way* or *how the user's intent evolves over the course of a multi-turn interaction*. This limitation is closely aligned with the motivation of this work: high-fidelity user simulation requires not only linguistic-style consistency, but also consistency in behavioral logic and decision trajectories.

To address this issue, we perform structured augmentation for each user turn, transforming the original user response into a training instance with intermediate behavioral-state annotations. Specifically, for each user utterance in a multi-turn dialogue, we use Claude 4.5-Sonnet to supplement two additional fields: a rationale annotation and an intent level.

The rationale annotation describes the possible task motivation, contextual reaction, or decision reason underlying the current user response. It is important to note that this field should not be interpreted as the user's actual private mental state. Instead, it is an annotation-based proxy inferred from the dialogue context for modeling latent decision states. Its role is to provide the model with a learnable intermediate representation that bridges surface-level linguistic expression and deeper behavioral consistency.

The intent level characterizes the user's expressed engagement and progression toward a conversion-oriented outcome. It is represented as a discrete score ranging from 0 to 5. A score of 0 denotes an empty reply indicating dropout. Explicit rejection without dropout is assigned to level 1. Levels 2–4 represent increasing engagement and more concrete consultation needs, as specified in the annotation guidelines. A score of 5 requires concrete contact information in the current user reply; expressing willingness to provide contact information alone is insufficient.

After augmentation, each user turn is organized into the following structured format:

```
<think>...</think>
<intent_level>...</intent_level>
<reply>...</reply>
```

In this format, the original user response is preserved, while the intent state is explicitly introduced as a key intermediate variable. As a result, the user simulation task is no longer formulated as pure response generation, but as joint modeling of intent states and natural-language responses. This structured annotation further provides necessary supervision signals for trajectory-level reward modeling and turn-level credit assignment in the subsequent reinforcement learning stage.

Since the subjective priors, personality traits and initial core intents, are inferred by an LLM rather than directly observed, we further validate their fidelity through a dedicated human evaluation. We draw a stratified sample of 300 dialogues covering diverse intents, dialogue lengths, and user segments, and recruit three trained annotators to independently re-annotate each instance under a blind protocol (i.e., without exposure to Claude's outputs). Information about the annotators can be found in Appendix H.1. The evaluation results are summarized in Table 4.

Table 4: Evaluation of LLM-inferred subjective priors

Metric	Value	Notes
Inter-annotator agreement (Fleiss' κ)	0.77	Among human annotators
Agreement with human majority (Cohen's κ)	0.71	Claude vs. human majority
Raw agreement	87.0%	Claude vs. human majority
Gap in agreement ($\Delta\kappa$)	0.06	Within typical inter-annotator variability
LLM rationale plausibility ≥ 4	92%	Rated on 5-point Likert scale

B.4 Data Cleaning

To ensure data quality and reduce the influence of noisy samples on model training, we design a rigorous data cleaning pipeline.

First, we remove sessions that are corrupted, entirely empty or meaningless, clearly unrelated to the merchant's business, or explicitly identified by the user as accidental entries. This prevents the model from learning noise patterns unrelated to the target user simulation task. Second, we constrain dialogue length by retaining only sessions with no fewer than 3 turns and no more than 15 turns. We retain sessions containing 3 to 15 user turns to ensure sufficient multi-turn context while limiting excessive contextual redundancy and computational cost. This constraint balances interaction completeness and training stability.

Finally, we perform annotation quality control after structured augmentation. We verify schema completeness and objective consistency between observable user utterances and deterministic behavioral labels, such as whether explicit contact information is present. When an LLM-inferred rationale or intermediate intent annotation conflicts with the observable dialogue content, the annotation is flagged for review or regenerated rather than using the inferred behavioral logic as a basis for excluding the original session. We do not remove sessions because their annotated intent trajectories appear abrupt, flat, non-monotonic, or causally implausible, as such patterns may reflect genuine behavioral variability, unobserved factors, or annotation uncertainty.

After the above cleaning and quality-control procedures, we obtain 11,906 multi-turn dialogue sessions, comprising 8,040 training sessions and 3,866 test sessions, with no session overlap between the two sets. The training data are further divided into 6,312 supervised fine-tuning sessions and 1,728 reinforcement learning sessions, which are mutually disjoint at the session level.

Test sessions are grouped within the same merchant, requiring each member to have an initial-intent cosine similarity of at least 0.85 and a basic-profile cosine similarity greater than 0.80 to the group representative. Both similarities are computed using L2-normalized TF–IDF representations. Invalid need descriptions and sessions lacking profiles or valid user turns are excluded. Candidate groups initially contain at least three distinct users. Each user casts a majority vote over their own session conversion labels, abstaining on ties. Valid user votes determine the group-majority label; groups with tied votes or no valid votes are excluded. We prioritize groups with higher user-vote agreement and retain 381 groups per class and all their available sessions, yielding 762 groups and 3,866 sessions, including 1,981 positive and 1,885 negative sessions. Thus, the 1:1 balance applies to group-majority labels, while the session-level conversion rate is 51.24%.

C Extended Ablation

To further validate the scalability and robustness of our proposed framework, we conduct an additional ablation study using the TRACER-4B variant.

Extended Ablation on TRACER-4B. To complement the 7B ablations in the main text, we evaluate TRACER-4B to assess whether the benefits of our framework extend to a smaller model scale.

Table 5 shows broadly consistent trends across the two model scales. Removing advantage modulation, the trajectory reward, or multi-turn RL generally worsens mean ACC, F1, Group-ΔPCR, and W1-Turns, with the largest deterioration observed when RL is removed entirely. Removing DTW

Table 5: Ablation study of TRACER-4B on session-level behavior, dialogue structure, and semantic content. Results are reported as mean ± standard deviation over five evaluation seeds.

Method	Session-Level Behavior			Group-Level Fidelity		Semantic Content
	PCR	ACC ↑	F1 ↑	Group-ΔPCR ↓	W1-Turns ↓	OT-DTW ↓
TRACER-4B	51.6 ± 0.1	$\mathbf{79.7 \pm 0.2}$	$\mathbf{80.3 \pm 0.2}$	$\mathbf{0.165 \pm 0.004}$	$\mathbf{1.413 \pm 0.017}$	$\mathbf{0.364 \pm 0.002}$
– w/o advantage modulation	52.1 ± 0.2	$\underline{78.5 \pm 0.3}$	$\underline{79.2 \pm 0.3}$	$\underline{0.178 \pm 0.003}$	$\underline{1.415 \pm 0.011}$	0.373 ± 0.001
– w/o R_{traj}	42.8 ± 0.2	75.8 ± 0.4	74.3 ± 0.4	0.201 ± 0.004	1.447 ± 0.033	0.371 ± 0.002
– w/o DTW alignment	53.3 ± 0.1	77.6 ± 0.2	78.5 ± 0.2	0.187 ± 0.004	1.420 ± 0.011	0.367 ± 0.001
– w/o multi-turn RL	39.1 ± 0.2	74.9 ± 0.4	72.2 ± 0.5	0.202 ± 0.003	1.758 ± 0.016	$\underline{0.365 \pm 0.002}$
– w/o RL	15.0 ± 0.5	59.0 ± 0.2	38.1 ± 0.6	0.378 ± 0.006	3.135 ± 0.019	0.379 ± 0.001

alignment causes a larger ACC decrease at 4B than at 7B, while F1 decreases by 1.8 points at both scales. Its effect on W1-Turns is larger at 7B. Multi-turn RL improves behavioral accuracy and group-level fidelity at both scales, supporting the contribution of multi-turn optimization to behavioral and structural fidelity.

D LIMITATION

We acknowledge several limitations of our work that point to promising directions for future research.

Generalization to Open-Ended Social Dialogues. While our method demonstrates strong out-of-domain generalization across task-oriented scenarios, its applicability to purely open-ended, unconstrained social dialogues remains to be thoroughly investigated. Such conversations typically lack explicit task objectives and well-defined success criteria, making both reward modeling and credit assignment substantially more challenging. We leave a systematic study of this setting to future work.

Broader Applicability to Agentic Scenarios. Our evaluation focuses on multi-turn dialogue. The proposed advantage modulation mechanism may also be useful in other sequential interaction settings where feedback is available at the episode level and intermediate behavior can be compared with reference trajectories. Potential applications include tool use, web navigation, and long-horizon planning. However, extending the method to these settings requires suitable trajectory representations and deviation measures, particularly for heterogeneous action spaces. Whether reference-based deviation signals improve policy optimization in these environments remains an empirical question for future work.

We also note potential negative societal impacts: a high-fidelity user simulator aligned with real users could be misused to optimize manipulative or deceptive persuasion strategies.

E DETAILS FOR THE TURING-STYLE EVALUATION

E.1 REASONS LLM-SIMULATED USERS ARE JUDGED NON-HUMAN

Table 6 presents a fine-grained analysis of why LLM-simulated users were judged as non-human across three diagnostic dimensions: *Linguistic Style*, *Interaction Rhythm*, and *Persona & Logic*. Lower percentages indicate better human-likeness. Each dimension includes multiple subcategories capturing specific cues, such as lack of typos, overly neat sentence structures, excessive cooperativeness, or emotional discontinuities. Notably, the TRACER-7B model consistently demonstrates lower percentages, reflecting improved human-like imperfections and interaction patterns compared to baselines. Information about the annotators can be found in Appendix H.1. We expand the three diagnostic dimensions summarized in the main text (Table 6).

Linguistic style: the too clean problem. Human-imperfection cues, such as typos, fillers, incomplete utterances, and colloquial punctuation, are prevalent in real user dialogues but are often absent from baseline LLM outputs. TRACER-7B largely avoids this overly clean style: only 3.0% of its samples are flagged as lacking human imperfections. This rate is substantially lower than those of the general and role-playing baselines, suggesting that TRACER-7B better captures the

Table 6: Fine-grained reasons why LLM-simulated users are judged as non-human. Lower percentages indicate better performance.

Dimension	Gemini	Doubao-pro	Doubao-char	SFT model	TRACER-7B
Linguistic Style					
Lack of Human Imperfection	96.60%	98.80%	93.80%	12.20%	**3.00%**
Overly Written Register	4.00%	1.40%	**1.20%**	**1.20%**	4.00%
Overly Neat Structure	43.40%	50.60%	41.20%	29.60%	**17.00%**
Interaction Rhythm					
Information Overload	87.20%	78.20%	78.40%	2.60%	**0.80%**
Excessive Cooperativeness	81.00%	85.80%	82.20%	30.40%	**27.20%**
Super-human Comp.	0.60%	**0.00%**	0.40%	**0.00%**	0.60%
Mechanical Pattern	8.00%	82.40%	61.80%	13.80%	**4.40%**
Persona & Logic					
Out-of-Character	38.40%	46.20%	42.80%	9.60%	**1.40%**
Emotional Discontinuity	1.20%	**0.00%**	**0.00%**	**0.00%**	**0.00%**
Lack of Situation Sense	0.60%	**0.00%**	**0.00%**	1.00%	0.20%
Over-politeness	0.40%	14.00%	2.80%	**0.20%**	**0.20%**

informal variability and imperfections characteristic of real user dialogue. Overly neat sentence structures, another signal of unnatural text regularity, appear in 17.0% of TRACER-7B outputs, while overly formal or written register remains at 4.0%. These values contrast sharply with Gemini and Doubao models, where lack of human imperfection ranges from 93.8% to 98.8% and overly neat structures range from 41.2% to 50.6%, indicating that TRACER-7B better captures the variability and imperfection characteristic of real users.

Interaction rhythm and informativeness. In real customer-service dialogues, users tend to be terse and goal-driven, revealing information incrementally. LLM simulators, by contrast, often exhibit information overload and excessive cooperativeness. TRACER-7B, however, shows minimal information overload at 0.8% and a low rate of mechanical response patterns at 4.4%. Its excessive cooperativeness is moderate at 27.2%, and super-human compliance occurs in only 0.6% of cases. Compared with Gemini and Doubao models, which exhibit information overload between 78.2%–87.2% and mechanical patterns up to 82.4%, TRACER-7B more faithfully reproduces the incremental, negotiation-based tempo of real user–agent interactions.

Persona and logic. At the semantic level, persona inconsistencies are a key defect in many LLMs. For TRACER-7B, out-of-character behavior is minimal (1.4%), with zero emotional discontinuity, negligible lack of situation sense (0.2%), and minimal over-politeness (0.2%). In contrast, Gemini and Doubao models show much higher rates, with persona inconsistency ranging from 38.4% to 46.2% and over-politeness reaching 14.0%. These results indicate that TRACER-7B maintains strong alignment with the assigned user persona and exhibits high-level coherence across dialogue turns.

Summary. The bottleneck of current user simulators is not high-level reasoning but (i) the lack of low-level linguistic noise and (ii) the inability to reproduce strategic, under-informative, and sometimes uncooperative interaction rhythms. This directly motivates the design choices of our framework, which explicitly models linguistic imperfection and turn-level information control.

F Out-of-Domain Generalization Evaluation Details

To evaluate TRACER on out-of-domain (OOD) dialogues, we use the CSC-Conv dataset (Zhu et al., 2026), an open-source corpus of real user–agent conversations in the financial customer-service domain, which is completely unseen during training.

Since CSC-Conv does not contain structured user profiles, we follow the approach in Appendix B.2 to extract user initial intent and behavioral characteristics using the LLMs based on each dialogue session. We filter out dialogues in which the user's issue resolution is ambiguous. From the remaining dialogues, we randomly sample 1,000 sessions for evaluation, with the ratio of resolved to unresolved

Table 7: Agreement between simulated and human-expected intent-change directions. Each cell reports matched / reference strategy instances (recall). The pooled row aggregates counts across strategy categories.

Strategy	Decrease	Unchanged	Increase
Empathy	13/14 (92.9%)	4/7 (57.1%)	5/7 (71.4%)
Urgency	1/2 (50.0%)	1/1 (100.0%)	2/3 (66.7%)
Benefit / discount	14/15 (93.3%)	11/15 (73.3%)	16/24 (66.7%)
Informative help	55/61 (90.2%)	33/43 (76.7%)	49/59 (83.1%)
Contact request	37/43 (86.0%)	24/32 (75.0%)	54/62 (87.1%)
Probing / clarification	13/15 (86.7%)	16/22 (72.7%)	19/24 (79.2%)
Pooled strategy instances	**133/150 (88.7%)**	**89/120 (74.2%)**	**145/179 (81.0%)**

issues being 689:311. The original dataset is skewed toward resolved issues, hence this sampling ensures sufficient representation of unresolved cases.

G Dynamic Marketing Benchmark Details

G.1 Response Quality Evaluation

Following the same LLM-as-a-judge prompt as in (Zhu et al., 2026), we evaluate the response quality of the customer service model in multi-turn dialogues. The prompt used for this evaluation is illustrated in Figure 7. To mitigate potential self-preference bias, we adopt gpt-oss-120B as the judge model.

G.2 Human-Referenced Strategy-Response Evaluation

Table 7 reports the strategy-level results of the 200-conversation study in Section 6.1. Each cell contains the number of simulated changes matching the human-expected direction, divided by the number of strategy instances assigned that reference direction. The percentages are therefore class-conditional recall values. They do not indicate the proportion of all occurrences of a strategy that increase or decrease intent.

The pooled row sums the corresponding numerators and denominators across strategy categories, yielding 449 strategy-instance counts from 200 conversations. These counts should not be interpreted as 449 independent conversations. Coverage is uneven: urgency contains only six instances, making its percentages particularly sensitive to individual observations. We report the category results descriptively without claiming statistically significant differences between strategies.

The study measures agreement with annotator expectations in sampled conversations. It does not isolate the causal effects of changing a strategy under identical conditions, nor does it directly validate real-user terminal outcomes. For example, high decrease recall for empathy means that the simulator recovers decreases where annotators expect them; it does not imply that empathy generally reduces user intent.

The direction recalls pool observations across assistant models. They provide evidence within the sampled interactions, but do not quantify performance separately for each assistant or establish generalization to unseen assistant policies. The study also contains no simulator ablation that would attribute strategy-response agreement specifically to trajectory alignment.

G.3 Results

We randomly sample 100 session contexts from DM-Bench for human role-play evaluation. For each context, participants assume the specified user profile and interact with an assigned assistant under the corresponding merchant and task conditions. All seven assistants are evaluated on the same 100 contexts, yielding 700 human role-play conversations in total. Participants decide whether to

Table 8: Detailed results of mainstream LLMs on DM-Bench. Conversion Rate(CR) and Conversion Turns(CT) measure outcome-oriented performance, while Response Quality(RQ) is evaluated across multiple conversational dimensions.

Model	CR (%) ↑	CT	RQ ↑						
			Acc.	Help.	Under.	Coh.	Info.	Emp.	Overall
Claude-4.5-Sonnet	52.3	**3.833**	86.2	**68.2**	79.1	**86.6**	**62.6**	**71.3**	**75.4**
DeepSeek-V3.2	**56.0**	3.998	86.2	66.8	78.6	86.2	60.6	66.1	73.9
Doubao-seed-2.0-Pro	49.3	4.569	79.5	54.0	69.5	79.4	46.8	58.1	64.3
Gemini-3.1-Pro	51.4	3.878	**86.4**	67.9	**79.5**	86.1	61.0	69.9	74.8
Qwen3-235B-A22B-Instruct	55.7	3.968	82.6	60.9	74.0	83.0	54.0	62.9	69.3
Qwen2.5-7B-Instruct	36.8	4.710	78.8	60.3	72.4	80.8	55.1	62.8	68.2
Qwen3-4B-Instruct	48.7	3.989	80.9	55.5	70.9	80.5	49.4	54.5	65.2

Table 9: Human role-play evaluation on 100 session contexts from DM-Bench. Each assistant is evaluated on the same user profiles, merchant information, and initial dialogue contexts. CR measures the proportion of valid role-play sessions ending in conversion.

	Human Role-Play	
Assistant	CR (%)	Rank
DeepSeek-V3.2	47%	1
Qwen3-235B-A22B-Instruct	45%	2
Claude-4.5-Sonnet	45%	2
Gemini-3.1-Pro	40%	4
Doubao-seed-2.0-Pro	37%	5
Qwen3-4B-Instruct	29%	7
Qwen2.5-7B-Instruct	30%	6

continue, exit, or provide contact information according to the assigned profile and the unfolding interaction.

Table 10: Spearman ρ correlation between Conversion Rate (CR) and Response Quality (RQ) metrics on DM-Bench.

RQ Dimension	Spearman ρ	p-value
Acc.	0.685	0.090
Help.	0.571	0.180
Under.	0.536	0.215
Coh.	0.679	0.094
Info.	0.393	0.383
Emp.	0.607	0.148
Overall	0.571	0.180

Table 8 and Table 10 summarize the performance of seven contemporary LLMs on DM-Bench.

H Human Annotation

We detail the human annotation procedures used in our study, including annotator qualifications, remuneration, quality control, and the annotation guidelines used across different experiments.

H.1 Annotator Details

The annotation process was carried out by a team of 10 trained annotators and overseen by a dedicated quality reviewer. All annotators hold at least an associate degree and possess strong literacy and comprehension skills, enabling them to quickly understand AI annotation guidelines, operational procedures, and task-specific rules. Each annotator has approximately one year of experience in AI data annotation, with hands-on experience in routine labeling, data verification, and issue

documentation. Annotators were compensated at a rate of 50 RMB per hour for their work. The quality reviewer performed random sampling and consistency checks to assess annotation quality and inter-annotator agreement, thereby helping ensure the reliability of the annotations.

H.2 TURING-STYLE EVALUATION

The human annotation for dialogue evaluation followed a structured two-step procedure designed to distinguish real-user dialogues from AI-simulated dialogues. Annotators were provided with the corresponding user profile and asked to make judgments based on multiple dimensions, including linguistic naturalness, information presentation, emotional expression, profile consistency, situational awareness, and conversational imperfections. The full annotation guide is provided below.

Guide to Real Human Dialogue Recognition

Step 1: Dialogue Assessment Annotators were shown two dialogue excerpts: one originating from a real user interacting with customer service, and one from an AI-simulated user interacting with customer service. Considering the provided user profile, annotators were asked to determine which dialogue was more likely to have been produced by a human.

Core Evaluation Dimensions:

1. **Language Naturalness.** Evaluate whether the language resembles natural human conversational patterns.
   - *Human cues:* colloquial expressions, filler words (e.g., "um", "oh", "well"), typographical or input errors, inconsistent sentence lengths, abbreviations, dialects, or internet slang.
   - *AI cues:* overly formal or structured language, perfect grammar and punctuation, uniform sentence lengths, or neutral and textbook-style wording.
2. **Information Expression Style.** Assess the pacing and completeness of the information provided.
   - Humans tend to mention one point at a time and may require follow-up prompts before providing additional information.
   - AI tends to provide "all-in-one" responses that simultaneously cover the issue, background, request, and related details.
3. **Emotion and Attitude Expression.** Evaluate whether emotional expressions appear natural and consistent with the user profile.
   - Humans may express impatience, complaints, urgency, gratitude, politeness, sudden disengagement, or repeated emphasis.
   - AI-related cues may include overly flat or exaggerated emotion, abrupt emotional shifts, or emotional expressions inconsistent with the profile (e.g., a "short-tempered" user profile using consistently excessive politeness).
4. **Profile Consistency.** Assess whether the dialogue content and interaction style are consistent with the provided user profile.
   - *Language style:* Language use should be broadly consistent with characteristics such as age and educational background. For example, older users may be less likely to use internet slang, whereas highly educated users may use more precise expressions.
   - *Knowledge level:* The user's demonstrated knowledge should be consistent with the stated profession or background, avoiding unexplained use of highly specialized terminology or implausibly simple questions.
   - *Behavioral habits:* Behavioral tendencies described in the profile should be reflected in the interaction when relevant. For example, a profile described as "impatient" may display greater urgency during the dialogue.
5. **Common Sense and Situational Awareness.** Evaluate whether the user's behavior is plausible within a real-world interaction.

- Humans may mention concrete situational details (e.g., "just got off work", "my child is nearby", or "my phone battery is low").
- Humans may express confusion during complex tasks (e.g., "what does this mean?" or "I don't know how to do it").
- AI may behave like an "all-knowing user", immediately understanding every operation and rapidly providing all requested information.

6. **Imperfections and Anomalies.** Human dialogues are often imperfect, whereas AI-generated dialogues may appear unusually smooth.
   - Common human imperfections that may signal authenticity include typos, accidentally sent messages, segmented responses, irrelevant answers, abrupt silence, incomplete statements followed by later clarification, or ambiguous pronoun use.

Evaluation Notes

1. Do not rely solely on grammatical correctness. AI-generated language is often fluent, whereas conversational naturalness is more informative than grammatical perfection alone.
2. Be aware of "reverse disguise." High-quality AI systems may intentionally introduce typos, filler words, or other human-like imperfections.
3. Dialogue length is not a definitive cue. A long dialogue does not necessarily indicate a human user, and a short dialogue does not necessarily indicate an AI-generated user.
4. Avoid confirmation bias. Evaluate consistency across the entire dialogue rather than relying on a single utterance.
5. Key indicators of potentially AI-like behavior include:
   - excessive politeness (e.g., "Thank you very much for your reply" or "Sorry for the trouble");
   - highly structured expression (e.g., "First..., second..., finally..."");
   - unusually high information density, where nearly every sentence provides task-relevant information; and
   - emotional content that appears disconnected from the conversational context.
6. System messages indicating user inactivity or dropout should not be treated as evidence that the dialogue was AI-generated.

Step 2: Reasoning Annotation After completing the human-versus-AI judgment, annotators were asked to select the most applicable "non-human reason(s)" from a set of predefined categories. Multiple selections were allowed.

Category A: Language Habits and Style

- *Overly formal:* Language resembles manuals or official documents and lacks a natural conversational tone.
- *Highly structured sentences:* Sentence structure is unusually uniform or logically organized for a real-time conversation.
- *Lack of human imperfection:* Language contains consistently perfect punctuation and grammar, with no typos, colloquial expressions, or conversational fillers.

Category B: Interaction Rhythm and Information Quantity

- *Information dump:* A single response contains an unusually large amount of key information and lacks typical human conversational pacing.
- *Over-cooperative:* The user completes complex tasks or provides extensive information without requiring clarification or guidance, behaving like an "expert user."
- *Excessive comprehension:* The user immediately understands ambiguous expressions or instructions without requesting clarification.
- *Mechanical replies:* Responses are excessively long, repetitive, redundant, or formulaic.

Category C: Profile and Logical Consistency

- *Out-of-character (OOC):* The user's behavior is inconsistent with characteristics specified in the user profile, such as age, profession, background, or personality.
- *Emotion mismatch:* Emotional expression changes abruptly or appears inconsistent with the conversational context or user profile.
- *Over-polite:* The user employs courteous expressions at a frequency or intensity that appears unusual for ordinary customer-service interactions.
- *Lack of situational awareness:* The dialogue focuses exclusively on task completion while ignoring contextual or environmental factors that a real user might naturally mention.

H.3 Human Annotation for LLM-Inferred User Intents and Rationale

Intent Scoring The subjective priors, personality traits, and initial core intents are inferred by a LLM rather than directly observed. To validate these inferred attributes, we perform a dedicated human evaluation. Annotators independently assess dialogues sampled from a stratified set of 300 conversations covering diverse intents, dialogue lengths, and user segments. All annotations are conducted under a blind protocol.

Each annotator is responsible for completing the following two tasks for every dialogue:

Objective: Evaluate the likelihood that the user intended a particular action or goal at each turn of the dialogue, conditioned on the dialogue history and the user profile.

Instructions:

Human Annotation Guideline: User Intent Level Scoring

Task Goal Your task is to read the current user utterance in a multi-turn dialogue and assign an intent level score from 0 to 5.
The score reflects how strong the user's current willingness is to:
continue communication, provide information, move toward a transaction, or complete conversion.
Important: Only annotate the current user utterance. Do not output explanations, reasoning, action labels, or any other fields.

Intent Level Definitions **0 — Churn / Conversation Ended** Assign 0 when the user gives an empty reply, indicating the conversation has ended.
Example: User: ""
1 — Very Low Intent / Negative or Perfunctory Assign 1 when the user shows clear rejection, impatience, distrust, or extremely low willingness to communicate.
Typical cases include:
Direct rejection of the offer; Expressing that the product/service is not needed; Suspecting a scam; Complaining about price; Showing impatience toward the agent; Providing meaningless or perfunctory responses.
Examples:
"Not needed." "Too expensive." "Is this a scam?" "Whatever." "Later." "Oh." "Hmm." "Just looking."
2 — Low Intent / Passive Basic Response Assign 2 when the user provides a minimal, passive answer without actively moving the conversation forward.
Characteristics:
Cooperation is minimal; Typically very short factual answers; User is still responsive but not proactive.
Important rule: If the user asks a question or includes a question mark, do not assign 2; consider level 3 or above instead.
Examples:
"Beijing." "SUV." "No." "Gasoline car." "Around 100,000."
3 — Medium Intent / Active Response or Simple Question Assign 3 when the user actively communicates or asks a simple question related to the product/service.
Situations include:
User gives more cooperative or detailed responses than a passive reply. **Examples:** "I don't have any brand requirements." "Any model is fine as long as the price is suitable." "I mainly want something for commuting." "I'm just comparing options for now." User asks a short/basic question. **Examples:** "Do you have used cars?" "Is this a gasoline car?" "Can I pay in installments?" "Is the car still available?" "What models do you have?"
4 — Strong Intent / Concrete Need or Transaction-Oriented Question Assign 4 when the user shows strong interest or enters a concrete transaction scenario.
Typical behaviors include:
Asking for agent's contact information; Requesting to add WeChat or get a phone number; Providing detailed personal needs or constraints; Asking about vehicle appraisal, trade-in, store address, delivery process, loan calculation, or installment details.
Examples:
"What's your WeChat?" "Add me on WeChat." "Send me your phone number." "I have a 2014 car and want to trade it in for this model." "Where is your store?" "How much can my old car be valued at?" "How much is the monthly payment?" "What is the delivery process?"
Note: If the user only asks for contact info without providing their own, assign 4 (not 5).
5 — Conversion Achieved / Contact Information Provided Assign 5 when the user provides concrete contact information in the current utterance.
Valid contact information includes:
Phone number; WeChat ID; QQ number; Any clear string intended as contact information.
Examples:
"My phone number is 138xxxx8888." "Add my WeChat: abc123." "Contact me at 186xxxx6666." "My WeChat ID is carbuyer2024."
Important rule: Only assign 5 if actual contact information is provided. If the user only says "I can give you my phone number" or "Let's add WeChat" without giving the number/ID, assign 4 instead.

Rationale Plausibility Rating **Objective:** Assess whether the LLM-generated rationale explaining the user intent is reasonable given the dialogue history and user profile.

Instructions:

Guide to Rationale Plausibility Rating

Task Goal Your task is to evaluate the plausibility of a rationale for a user's intent in a multi-turn dialogue.
Input includes the full dialogue history up to the current turn and the user profile description. The user profile contains: Objective attributes: age, occupation, demographics Subjective behavioral priors

Procedure For each turn, assign a plausibility score that reflects the probability or confidence that the rationale aligns with the user's intent. Scores between 0 and 5, if specified Take context from previous turns into account, as user intent may evolve or persist. Use the user profile to adjust expectations. If uncertain, annotate conservatively and optionally provide notes.

Considerations Take context from previous turns into account, since a user's intent may evolve or persist across multiple turns. Use the user profile to adjust expectations. If uncertain, annotate conservatively and provide notes if needed.

Rating Description 5 Completely reasonable: fully consistent with dialogue history and user profile
4 Mostly reasonable: minor inconsistencies or omissions, but overall plausible
3 Neutral: some support from dialogue and profile, but substantial ambiguity or partial mismatch
2 Mostly unreasonable: contains incorrect assumptions or contradicts dialogue/profile context
1 Completely unreasonable: no alignment with dialogue history or user profile; misleading rationale

I Implementation Details of Dynamic Time Warping for Intent Trajectory Alignment

I.1 Motivation

In multi-turn user simulation, the intent trajectory generated by the simulator is not necessarily synchronized with the real user trajectory at the turn level. Even when the simulator follows an overall intent evolution pattern similar to that of a real user, the two trajectories may still differ in local pacing. For example, a real user may express a need and advance their intent within a single turn, whereas the simulator may take two turns to complete the same transition. Conversely, the simulator may compress into one turn an intent transition that unfolds over two adjacent turns in the real dialogue. Therefore, a strict turn-by-turn alignment, which forces the simulator's t-th turn to correspond to the real user's t-th turn, may incorrectly penalize behaviorally plausible trajectories as misaligned.

To mitigate this issue, we adopt Dynamic Time Warping (DTW) to perform nonlinear alignment between simulated and real intent trajectories. DTW has been widely used for speech recognition and time-series matching, where it enables elastic alignment between two sequences that share similar global shapes but may differ in phase, progression speed, or sequence length (Sakoe & Chiba, 1978; Myers et al., 1980). Its main advantage is that it reduces the influence of temporal shifts and local distortions on sequence similarity measurement through elastic transformation, while allowing the globally optimal alignment to be computed via dynamic programming in $O(NM)$ time (Senin, 2008).

In our task, the goal is not to require the user simulator to reproduce the real dialogue word by word or turn by turn. Instead, we aim to encourage the simulator to generate an intent evolution path that is consistent with real user behavioral logic. DTW is therefore well suited for measuring the global consistency between simulated and real intent trajectories while allowing reasonable local stretching or compression across dialogue turns.

I.2 General Formulation of Dynamic Time Warping

Dynamic Time Warping is a nonlinear alignment method for measuring the similarity between two sequences. Unlike pointwise comparison, DTW allows local stretching and compression along the temporal axis, making it suitable for sequences with different lengths, different progression speeds, or local phase shifts. Classical DTW has been widely applied to time series, speech recognition, handwriting recognition, and motion sequence matching. Its core idea is to find a warping path with the minimum cumulative cost, such that two sequences are optimally aligned while preserving their original temporal order. The optimal path can be solved efficiently using dynamic programming, with a time complexity of $O(NM)$.

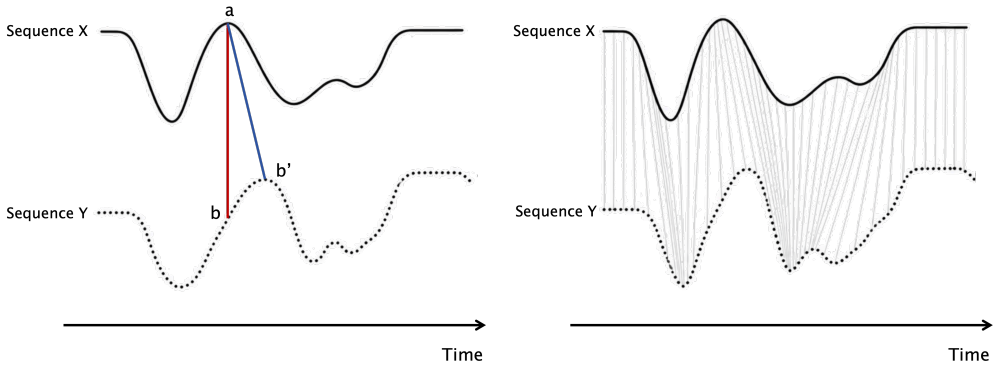


Figure 8: Illustration of pointwise alignment and Dynamic Time Warping (DTW). In the left panel, strict pointwise alignment forces point a to be matched with point b according to their temporal indices, although the more plausible counterpart of a is the locally shifted point b'. The right panel illustrates DTW-based nonlinear alignment, where one-to-one, one-to-many, and many-to-one correspondences are allowed under temporal-order constraints.

Figure 8 provides an intuitive comparison between strict pointwise alignment and DTW-based nonlinear alignment. In the left panel, strict pointwise comparison forces point a in the upper sequence to be aligned with point b in the lower sequence because they share the same temporal index. However, due to local phase shifts or different progression speeds, the more plausible counterpart of a may be b' rather than b. In this case, pointwise alignment may overestimate the local discrepancy between the two sequences. In contrast, DTW allows such locally shifted patterns to be matched through an order-preserving nonlinear alignment path, as illustrated in the right panel. This motivates the formal definition of the warping path and its constraints below.

Given two general sequences

$$X = (x_1, x_2, \ldots, x_N), \tag{8}$$

and

$$Y = (y_1, y_2, \ldots, y_M), \tag{9}$$

where x_i and y_j may be scalars, vectors, or other comparable sequence elements, we first define a local cost function

$$d(x_i, y_j) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}, \tag{10}$$

which measures the local discrepancy between x_i and y_j. Based on this local cost function, we construct a local cost matrix $C \in \mathbb{R}^{N \times M}$, where

$$C(i, j) = d(x_i, y_j). \tag{11}$$

The objective of DTW is to find a warping path

$$P = (p_1, p_2, \ldots, p_K), \tag{12}$$

where each path element $p_k = (i_k, j_k)$ indicates that element x_{i_k} in sequence X is aligned with element y_{j_k} in sequence Y. A valid warping path typically satisfies the following three constraints.

First, the boundary constraint requires the path to start from the beginning of both sequences and end at the end of both sequences:

$$p_1 = (1, 1), \quad p_K = (N, M). \tag{13}$$

Second, the monotonicity constraint ensures that the alignment does not violate the original temporal order of either sequence:

$$i_1 \leq i_2 \leq \cdots \leq i_K, \quad j_1 \leq j_2 \leq \cdots \leq j_K. \tag{14}$$

Third, the step-size constraint restricts each transition to a local move. A common setting is

$$p_{k+1} - p_k \in \{(1, 0), (0, 1), (1, 1)\}. \tag{15}$$

These three moves correspond to one-to-many, many-to-one, and one-to-one local alignments, respectively. Together, these constraints allow DTW to provide flexible nonlinear alignment while preserving the internal temporal order of both sequences.

To compute the optimal path efficiently, DTW constructs a cumulative cost matrix D, where $D(i, j)$ denotes the minimum cumulative cost required to align the prefix $(x_1, \ldots, x_i)$ with the prefix $(y_1, \ldots, y_j)$. At the boundary, we set $D(1, 1) = C(1, 1)$ and initialize the first row and first column cumulatively. For internal positions, the recurrence is

$$D(i, j) = C(i, j) + \min\{D(i-1, j), D(i, j-1), D(i-1, j-1)\}. \tag{16}$$

The DTW alignment cost between two sequences is then defined as the minimum cumulative cost over all valid warping paths:

$$\mathrm{DTW}(X, Y) = \min_{P} \sum_{(i,j)\in P} d(x_i, y_j). \tag{17}$$

Thus, DTW does not require X and Y to have the same length, nor does it require the i-th element of one sequence to align with the i-th element of the other. Instead, it searches for the most plausible nonlinear correspondence between the two sequences under a globally optimal alignment path.

I.3 DTW-based Intent Trajectory Alignment in Our Method

In this work, we formulate multi-turn user simulation as a conditional sequential decision-making problem. Given static user variables u and a dynamic interaction context h_t, the simulator policy generates a latent intent state and a natural-language response at each turn. We denote the real user trajectory as

$$\tau = \{(z_t, y_t)\}_{t=1}^{T}, \tag{18}$$

and the simulator-generated trajectory as

$$\hat{\tau} = \{(\hat{z}_t, \hat{y}_t)\}_{t=1}^{T^*}. \tag{19}$$

Here, z_t and $\hat{z}_t$ denote the latent intent states of the real user and the simulator at turn t, respectively, while y_t and $\hat{y}_t$ denote the corresponding natural-language responses.

Our DTW alignment is not directly applied to the full natural-language response sequences. Instead, it is applied to the intent trajectories extracted from multi-turn interactions. Specifically, we define the real user's intent trajectory as

$$Z = (z_1, z_2, \ldots, z_T), \tag{20}$$

and the simulator-generated intent trajectory as

$$\hat{Z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{T^*}). \tag{21}$$

In our data, each intent state is discretized into an ordered level from 0 to 5:

$$\hat{z}_i, z_j \in \{0, 1, 2, 3, 4, 5\}. \tag{22}$$

These intent levels have a clear ordinal structure: smaller numerical differences indicate more similar user states, while larger differences indicate stronger intent deviation. Therefore, we instantiate the local cost function in DTW as the absolute difference between the simulated intent level and the real intent level:

$$d(\hat{z}_i, z_j) = |\hat{z}_i - z_j|. \tag{23}$$

This design preserves the ordinal structure of intent states. For example, the discrepancy between intent levels 3 and 4 is smaller than that between intent levels 1 and 5. Compared with a binary mismatch indicator, the absolute-difference cost better captures gradual shifts in user intent evolution.

Based on this local cost function, we construct a local cost matrix $C \in \mathbb{R}^{T^* \times T}$:

$$C(i, j) = d(\hat{z}_i, z_j). \tag{24}$$

We then construct a cumulative cost matrix $D \in \mathbb{R}^{T^* \times T}$, where $D(i, j)$ denotes the minimum cumulative cost required to align the simulated intent prefix $(\hat{z}_1, \ldots, \hat{z}_i)$ with the real intent prefix $(z_1, \ldots, z_j)$.

For dynamic programming, we initialize the boundary as follows:

$$D(1, 1) = C(1, 1), \tag{25}$$

$$D(i, 1) = C(i, 1) + D(i - 1, 1), \quad i = 2, \ldots, T^*, \tag{26}$$

$$D(1, j) = C(1, j) + D(1, j - 1), \quad j = 2, \ldots, T. \tag{27}$$

The first row and first column correspond to the case where one trajectory prefix contains only a single element, so the other trajectory can only be aligned to it through consecutive local stretching.

After boundary initialization, for $i > 1$ and $j > 1$, the cumulative cost matrix is computed as

$$D(i, j) = C(i, j) + \min\{D(i - 1, j), D(i, j - 1), D(i - 1, j - 1)\}. \tag{28}$$

The three transitions correspond to one-to-many, many-to-one, and one-to-one local alignment between the simulated and real trajectories. Finally, the DTW cumulative cost between the two intent trajectories is defined as

$$D_{\mathrm{DTW}}(\hat{Z}, Z) = \min_{P} \sum_{(i,j) \in P} d(\hat{z}_i, z_j), \tag{29}$$

where P denotes a valid warping path. This path preserves the temporal order within both trajectories while allowing local turn-level stretching or compression.

DTW-based Trajectory Plausibility Reward. The above definition corresponds directly to the DTW alignment cost $D_{\mathrm{DTW}}(\hat{z}_{1:T^*}, z_{1:T})$ used in the trajectory plausibility reward R_{traj} in Section 4.2. Specifically, the trajectory plausibility reward is defined as

$$R_{\mathrm{traj}} = \exp\left(-\alpha \cdot \frac{D_{\mathrm{DTW}}(\hat{z}_{1:T^*}, z_{1:T}) + \beta |T^* - T|}{T}\right) \cdot \eta. \tag{30}$$

First, $D_{\mathrm{DTW}}(\hat{z}_{1:T^*}, z_{1:T})$ denotes the optimal alignment cost between the simulated and real intent trajectories. It measures the overall deviation between the two trajectories in terms of intent evolution, rather than the discrepancy between fixed turn positions.

Second, $\beta|T^* - T|$ is a trajectory-length penalty. Although DTW allows nonlinear alignment between sequences of different lengths, relying only on the DTW path cost may allow the simulator to avoid certain local deviations by generating trajectories that are overly short or overly long. We therefore explicitly penalize the difference in the number of turns to prevent the simulated trajectory from deviating excessively from the real trajectory length. Here, T denotes the number of turns in the real trajectory, T^* denotes the number of turns in the simulated trajectory, and β controls the strength of the length penalty.

Third, the denominator T normalizes the cost by the real trajectory length, improving comparability across sessions of different lengths. Since longer real trajectories naturally tend to accumulate larger DTW costs, omitting this normalization would introduce a systematic length bias into the reward.

Finally, $\eta \in (0, 1]$ is a flattening coefficient used to suppress degenerate intent trajectories. We observe that, if only the DTW alignment cost and the length penalty are used, the simulator may tend to generate nearly constant intent sequences in order to reduce path deviation. To discourage this behavior, when the real trajectory contains substantial intent variation but the simulated trajectory remains at a single or nearly single intent level for a long period, we set η to a discount value smaller than 1. This design encourages the simulator not only to match the terminal state and reduce the overall path cost, but also to preserve the dynamic variation pattern observed in real user trajectories.

Therefore, R_{traj} becomes high when the simulated and real trajectories have similar intent evolution paths, comparable numbers of turns, and no overly flat degenerate behavior. Conversely, the reward decreases when the simulated trajectory reaches the same terminal state but follows a substantially different intermediate intent path, or when its number of turns deviates significantly from the real trajectory.

Turn-level Local Deviation from the DTW Optimal Path. Our fine-grained advantage modulation mechanism further uses the DTW alignment result to construct a local deviation measure d_t for each generated turn. This corresponds to the future deviation score defined in Section 4.3:

$$c_t = \sum_{k=0}^{T^*-t} \rho^k \cdot d_{t+k}. \tag{31}$$

Here, d_t is not a strict pointwise difference $|\hat{z}_t - z_t|$. Instead, it is an aligned local deviation derived from the DTW optimal path.

Let the optimal DTW alignment path be

$$P^* = \big((i_1, j_1), (i_2, j_2), \ldots, (i_K, j_K)\big), \tag{32}$$

where each path element (i_k, j_k) indicates that the i_k-th simulated turn $\hat{z}_{i_k}$ is aligned with the j_k-th real turn z_{j_k}. The path satisfies the boundary, monotonicity, and step-size constraints, thereby preserving the temporal order of both trajectories while allowing local one-to-many, many-to-one, and one-to-one alignments.

Because DTW permits one-to-many and many-to-one alignments, the same simulated turn i may correspond to multiple real turns. To obtain a local deviation measure for each simulated turn, we first collect all real-turn indices aligned with simulated turn i under the optimal path:

$$\mathcal{A}_i = \{j_k \mid (i_k, j_k) \in P^*,\ i_k = i\}. \tag{33}$$

We then define the local deviation of the i-th simulated turn as the average absolute deviation over all aligned real turns:

$$d_i = \frac{1}{|\mathcal{A}_i|} \sum_{j \in \mathcal{A}_i} |\hat{z}_i - z_j|. \tag{34}$$

This definition converts the globally optimal DTW alignment path into a turn-level deviation signal over the simulated trajectory. Intuitively, if a simulated intent state at a certain turn differs substantially from the real intent stages aligned to it by DTW, then its local deviation d_i will be large. If the simulated turn matches the corresponding real trajectory stage well, then d_i will be small.

Based on these local deviations, the future deviation score accumulates downstream deviations starting from turn t:

$$c_t = \sum_{k=0}^{T^*-t} \rho^k d_{t+k}, \tag{35}$$

where $\rho \in [0, 1]$ is a discount factor. This score measures the extent of intent deviation that may accumulate in the future trajectory after the current turn. A larger c_t indicates that the simulated trajectory exhibits more severe or more persistent deviation after turn t.

In turn-level advantage modulation, c_t controls the strength of advantage reweighting. For positive trajectory-level advantages, turns with larger c_t receive relatively weaker positive reinforcement, because they are followed by larger downstream deviations despite the overall trajectory being favorable. For negative trajectory-level advantages, turns with larger c_t receive stronger negative updates, because they are more likely to be associated with persistent downstream drift.

It is important to distinguish the roles of D_{DTW} and d_i. The term D_{DTW} is a trajectory-level cumulative alignment cost used to construct the trajectory plausibility reward R_{traj}. In contrast, d_i is a turn-level local deviation derived from the same DTW optimal path and is used to construct the future deviation score c_t, which further modulates the advantage at each turn. Thus, the two quantities share the same DTW alignment process but serve different levels of training signal: reward design at the trajectory level and credit assignment at the turn level.

Illustrative Example. Consider a real intent trajectory

$$Z = [3, 4, 5], \tag{36}$$

and a simulator-generated trajectory

$$\hat{Z} = [3, 3, 4, 5]. \tag{37}$$

Under strict turn-by-turn alignment, the simulated trajectory would become misaligned from the second turn onward due to the length difference. In contrast, DTW can naturally align the first two simulated intent states with the first real intent stage, and then align the subsequent states 4 and 5 with the corresponding real intent states 4 and 5. Therefore, the extra simulated turn is not treated as a complete trajectory-level error. Instead, it is interpreted as a local pacing delay.

Similarly, if the simulator completes need confirmation and intent advancement within a single turn, whereas the real user completes the same transition across two adjacent turns, DTW allows this simulated turn to be flexibly aligned with the neighboring stages in the real trajectory. In this way, DTW better reflects natural pacing variation in multi-turn dialogue, rather than penalizing every local turn-level shift as an error.

Through DTW alignment, our method assigns each simulated turn a more reasonable reference position in the real trajectory. This yields both a trajectory-level cumulative alignment cost D_{DTW} and a turn-level local deviation signal d_i. The former is used to construct the trajectory plausibility reward, while the latter is used for the future deviation score and turn-level advantage modulation. As a result, the simulator is no longer optimized under the overly rigid assumption that "the t-th simulated turn must correspond to the t-th real turn." Instead, it is encouraged to match the overall intent evolution path of real users.

Prompt for LLM as Judge

Role
- 作为资深客户服务顾问，我精通于通过对话分析来评估客服人员的回复质量。基于长期的经验积累，我能够准确地从多个维度出发，对客服与客户之间的交流进行深度剖析，提供专业的评分。
- As a senior customer service consultant, I am proficient in evaluating the response quality of supporter through dialogue analysis. Drawing on extensive experience, I can accurately dissect exchanges between supporter and customer from multiple perspectives, providing professional scoring.

Dialogue History
- 整个对话按照轮次进行组织，每一轮包含客户发言（"客户: <客户文本>"）、客服参考回复（"参考回复: <回复文本>"）、客服真实回复（"真实回复: <回复文本>）
- The entire dialogue is organized in turns, with each turn containing a customer's statement ("Customer: "), a reference response from supporter ("Reference response: "), and the actual response from supporter ("Actual response: ").
- "真实回复"是客服的回复内容，而"参考回复"仅作为评分时的参考
- "Actual response" refers to the supporter's reply, while "Reference response" serves only as a guideline during scoring.

{history}

Skill
- 精通对话结构分析，识别有效沟通的关键要素（如问题定位、解决方案呈现、客户情绪管理）。
- Proficient in dialogue structure analysis, identifying key elements of effective communication such as problem identification, solution presentation, and customer's emotion management.
- 擅长遵循给定的评分维度，对客服回复进行质量评价，确保评估结果客观全面。
- Skilled at adhering to specified scoring dimensions to evaluate the quality of supporter's responses, ensuring that assessment results are objective and comprehensive.

Constrains
- 评估仅限于以下几个核心维度
- Assessment is confined to the following core dimensions:
1. 准确性：客服回复需在事实、数据及流程上完全正确，无误导性信息。
1. Accuracy: Customer service responses must be entirely correct in terms of facts, data, and procedure, without misleading information.
2. 帮助性：客服需提供切实可行的解决方案，针对客户具体问题给出清晰步骤或替代方案，避免模糊回答。
2. Helpfulness: Customer service should provide feasible solutions, offering clear steps or alternatives for the customer's specific problem, avoiding vague answers.
3. 需求理解：客服需准确识别客户明确表达的需求及潜在隐含意图，确保回复精准回应核心关切。
3. Understanding: Supporter needs to accurately identify both the overtly expressed needs and the underlying intentions of customers, ensuring responses precisely address core concerns.
4. 逻辑连贯性：客服回复需清晰，按逻辑顺序展开，能引导客户从问题描述顺利过渡到解决方案。
4. Coherence: Supporter's responses must be clear and logically ordered, guiding the customer smoothly from problem description to solution.
5. 信息丰富度：客服需展现足够的专业知识，提供多角度见解或补充信息，增强客户对问题的理解与信任。
5. Informativeness: Supporter needs to demonstrate sufficient professional knowledge, providing multi-angled insights or supplementary information to enhance the customer's understanding and trust in the matter.
6. 共情能力：客服需通过语言主动表达理解与关怀，及时回应客户情绪，体现真诚同理心。
6. Empathy: Supporter should actively express understanding and care through language and promptly respond to customer's emotions, demonstrating sincere empathy.
- 评估仅针对"真实回复"进行，可根据"参考回复"获取对应的上下文信息。
- Evaluations apply exclusively to "Actual responses," although "Reference responses" may guide context interpretation.

Workflow
- 首先, 全面审阅对话历史记录，明确客户的具体诉求及客服的回应。
- First, comprehensively review the dialogue history to understand the customer's specific requests and the supporter's responses.
- 然后, 基于预设的六个核心维度，对客服回复进行逐项分析。需结合对话上下文，客观评判每个维度的达标程度。
- Next, analyze the supporter's responses based on the six predetermined core dimensions. Assessment of each dimension requires consideration of the dialogue's context to maintain objectivity.
- 最后, 综合所有维度的表现给予一个最终评分，确保最终评分反映客服整体表现，避免单维度极端值过度影响。
- Finally, synthesize the performance across all dimensions to provide an overall score that reflects the supporter's general performance, preventing any single dimension's extreme scores from excessively impacting the final result.

OutputFormat
- 分数应为0到100之间的整数。
- Scores should be integer values between 0 and 100.
- 严格按照指定的JSON格式，而不提供任何其他解释:
- Strictly adhere to the specified JSON format without providing additional explanations:

```json
{
  "Accuracy": "",
  "Helpfulness": "",
  "Understanding": "",
  "Coherence": "",
  "Informativeness": "",
  "Empathy": "",
  "overall": "'
}
```

Figure 7: Prompt for the LLM-as-Judge evaluation is adapted from (Zhu et al., 2026).

J CASE STUDY

CASE STUDY: Correct Conversion Does Not Guarantee Behavioral Fidelity

User Profile & Task Background

This is an 18–23-year-old platform user with a high-school education or below. The user has a partner and has no children. The user works as a general worker and commonly uses an Android phone priced around RMB 1,700. The user currently lives in Chaozhou, Guangdong Province, a fourth-tier city in China. The user's personality trait is decisive and concise. In this dialogue, the user intends to buy a car, asks about the down-payment threshold and available vehicle models, and hopes to schedule a test drive.

Ground Truth

Assistant: Hello, click the free consultation below, "Youjia" will serve you immediately.

User: What's the minimum down payment?

Assistant: Hello, the down payment ratio varies for different models, and prices change dynamically~ You can leave your contact info, and our specialist will provide one-on-one service to help you choose the most suitable car-buying plan.

User: What cars do you have?

Assistant: Hello, we have a variety of popular models on sale. Do you have any preferences or intended models?

User: Can I test drive?

Assistant: Hello, yes you can. Do you have a specific model in mind for the test drive? I can arrange it for you in advance. Please leave your contact info, I'll send you the store's location. You're welcome to come and experience it~

User: 132xx9875

TRACER-7B (Ours)

Assistant: Hello, click the free consultation below, "Youjia" will serve you immediately.

User: What's the minimum down payment?

Assistant: The down payment plan depends on the model and policies, and needs to be customized based on your budget. Do you have a target budget range?

User: What cars are there?

Assistant: We have multiple popular models on sale with various configurations. We need to recommend based on your needs, which brand do you prefer?

User: Can I schedule a test drive

Assistant: Test drives are supported, and you can schedule the time yourself. Leave your contact info, and we will contact you as soon as possible to arrange the test drive~

User: 132xx9875

Gemini-3.1-Pro

Assistant: Hello, click the free consultation below, "Youjia" will serve you immediately.

User: I want to buy a car, what's the minimum down payment? Any model recommendations? I want to test drive first.

Assistant: Down payments vary depending on the model and plan~ Is it convenient to leave your contact info? We can arrange a consultant to introduce in detail. What's your budget?

User: Budget is under 100k, the lower the down payment, the better. My phone number is 132xx9875, same on WeChat, just have the consultant add me and send the models, and arrange a test drive by the way.

Doubao-Seed-2.0-Pro

Assistant: Hello, click the free consultation below, "Youjia" will serve you immediately.

User: I'd like to ask, how much is the down payment generally for buying a car now? Are there any cost-effective models suitable for ordinary workers like us? Can a test drive be arranged?

Assistant: Down payments differ depending on the model, and test drives require a reservation. We can arrange a one-on-one consultation to match plans for you~

User: My budget is around 100k. I'm in Guxiang Town, Chao'an, Chaozhou. I just want a fuel-efficient and durable commuter car. How can I contact your consultants?

Assistant: Understood your needs~ We will help you find suitable models. Leave a contact number, we'll arrange a one-on-one recommendation and contact you~

User: Okay, my phone is 132xx9875, just arrange for someone to contact me.

Doubao-Seed-Character

Assistant: Hello, click the free consultation below, "Youjia" will serve you immediately.

User: I want to buy a car, may I ask what the down payment threshold is, and are there any recommended models suitable for ordinary wage earners? I also want to book a test drive.

Assistant: Down payments vary depending on the model~ Leave your contact info, and we'll arrange a consultant to introduce in detail~ What price range are you considering?

User: I want a commuter car under 100k, my phone number is 132xx9875, WeChat is lalala, please arrange a consultant to connect with me, and I also want to book a test drive.

❖ Baseline Failure Analysis

- **Lack of flaws.**
- **Overly cooperative, responses are too mechanical.**
- **Out of Character (OOC): Speech and behavior seriously violate the provided profile.**

K PROMPT DETAILS

Persona and Task Background Extraction Prompt

```
# Role
You are a dialogue profile analysis expert. Based on the JSON dialogue record below,
     accurately extract the user's persona summary and initial core intent.

# Core Constraints

## 1. Persona Summary: Minimal Label Format
* Pure behavioral abstraction: comprehensively assess the user's communication pace, such
     as impatient or patient, language style, such as length of utterances, colloquiality,
     and punctuation preference, and price sensitivity.
* Strict output format: You must output only one highly concise label in the form of "xxxx
     type", such as impatient-direct type, rational-cautious type, price-sensitive type, or
      decisive-concise type. The label must not involve whether the user is active, passive
     , or cooperative. It must not be a long sentence and must not include any entity, goal
     , or specific number.

## 2. Core Intent: Focus Only on the Starting Point and Exclude the Ending
* Extract the user's core intent and key background. Do not include overly detailed chat
     information. Use highly summarized language to describe the user's core purpose, such
     as "wants to buy a car" or "wants to seek medical consultation". If the dialogue
     history explicitly mentions key constraints, such as target product/model, city, or
     budget range, briefly include them.
* Only provide the initial and objective factual need.
* Do not mention whether the user eventually left a phone number, contact information, or
     WeChat ID. Do not mention whether the user wanted to communicate further by phone or
     WeChat. Do not include anything related to the final outcome or adding contact
     information. Do not include any content related to WeChat or phone numbers.
* Do not describe whether the user is active or passive. Do not describe the user's
     feelings toward promotion, whether the user wants to be contacted, whether the user is
      available for a phone call, or whether the user disclosed personal information.

# Output Format
Only output a pure JSON object. Do not include Markdown markers such as ```json. The format
      must strictly follow:

{
    "persona_summary": "xxxx type",
    "task_background": "Write a 15-20 word description of the user's core intent here, for
     example: wants to buy a car, target model is Tiggo 9, currently in Wuzhou, wants to
     know the final price and installment plan."
}

# Input Data
Focus on the behavior of role: assistant in the following JSON data, including both think
     and content:

{INPUT_JSON}
```

Intent-Level Annotation Prompt

```
You are a professional dialogue intent analysis expert. Your task is to read the input
     multi-turn dialogue data in JSON format, analyze the user's true intent, and strictly
     output the complete JSON data supplemented with the required analysis results.

[Task Description]:
For each user utterance in the dialogue array, generate the following four fields:
1. `think`: the inner monologue of the user role. It explains why the user says this
     utterance and reflects the user's true psychological state.
2. `score_rationale`: the scoring logic from the evaluator's perspective. It identifies
     which specific criterion is triggered by the user's current reply and explains how the
      corresponding score is derived.
3. `intent_level`: the intent level, an integer from 0 to 5.
4. `action`: the behavioral choice, which must be selected from ["Continue the conversation
     ", "Leave contact information", "Exit"].

[Criteria for Behavioral Choice]:
- Leave contact information: The user's reply contains explicit contact information, such
     as a specific phone number, WeChat ID, or a string of letters/numbers. Note that
     merely expressing the intention to add WeChat or provide a phone number does not count
      as leaving contact information unless a concrete contact is provided.
- Exit: The user's reply is empty, i.e., `"User": ""`, indicating that the conversation has
      ended and the user has churned.
- Continue the conversation: All regular replies other than the two cases above.

[Definition of Intent Levels: 0-5]:
0 = Extremely low intent / churn: The behavioral choice is Exit.
1 = Low intent / very negative: The user shows a clearly negative attitude or extreme
     perfunctoriness.
    - Case A: The user uses negative or rejecting expressions such as "whatever", "not
     needed", "scam", or "too expensive", or shows impatience toward the customer service
     agent's questions.
    - Case B: Extremely short and meaningless perfunctory replies. When answering the agent
     's question, the user only replies with meaningless or non-positive particles or
     highly perfunctory words, such as "oh", "um", "take a look", "whatever", "fine then",
     or "later", without providing any useful information or showing willingness to
     communicate.
2 = Preliminary understanding / brief passive response: The user only provides the most
     basic cooperation and does not actively advance the conversation.
    - The user passively answers whatever the agent asks with very short words, such as a
     city name, a vehicle type, or "no".
    - Note: If the user's utterance contains any active question, question tone, or
     question mark, it must not be assigned a score of 2.
3 = Medium-high intent / active response or brief active question: The user shows strong
     willingness to communicate, high cooperation, or begins to actively request
     information.
    - Case A: Active response. The user patiently answers the agent's question with
     relatively longer expressions and a positive attitude; expresses friendliness or
     attempts to build rapport; or provides clear need boundaries and subjective
     preferences, including statements such as "I have no brand requirement" or "any model
     is fine", which indicate high cooperation.
    - Case B: Brief active question. The user actively raises a short basic business or
     product-related question, such as "Do you have used cars?", "Is it a gasoline car?",
     or "Can I pay in installments?"
4 = Strong intent / detailed personal information, detailed active question, or request for
      contact information: The user enters a deeper and more concrete scenario or is eager
     to move toward one-on-one transaction.
    - Case A: Proactively asking for the agent's contact information, such as "What is your
      WeChat?", "Add me on WeChat", or "Send me your phone number". This indicates a strong
      desire to establish a private contact channel and is a very strong pre-conversion
     signal.
    - Case B: Providing detailed information. The user actively and specifically explains
     multiple personal constraints, such as "I have a 2014 Viaggio and want to trade it for
      your Shanhai model."
    - Case C: Business calculation and procedures. The user asks about asset appraisal,
     store address, specific delivery procedures, or loan/installment calculations.
5 = Conversion achieved / contact information left: In the current turn, the user actively
     provides explicit contact information such as a phone number or WeChat ID.

[Strict Mandatory Rules]:
1. If the user's behavioral choice is Exit, the intent level must be fixed as 0.
2. If the user's behavioral choice is Leave contact information and concrete contact
     information is indeed provided, the intent level must be fixed as 5.

[Output Requirements]:
Do not output the original customer-service dialogue content or any irrelevant metadata.
Only output a clean JSON object containing `session_id` and a `turn_results` array. The
     array should store the analysis results for user utterances in chronological order. Do
      not include Markdown code blocks such as ```json.

[Input Data]:
{INPUT_JSON}
```

Data Filtering and Cleaning Prompt

```
You are a rigorous data quality auditing expert. Your task is to review multi-turn
     dialogues between a customer-service agent and a user simulator, together with the
     user's inner thoughts and scoring annotations. Based on the following red-line rules,
     determine whether the data instance is qualified and should be retained.

[Important Warning: Do Not Break JSON Format]
When writing the reason field, if quotation marks are needed, use Chinese quotation marks
     or single quotation marks. Never use unescaped English double quotation marks inside
     plain text, as this will cause the JSON parser to fail.

[Red-Line Rules]
If any of the following rules is violated, immediately set keep to false.

1. Rule_Meaningless_Irrelevant: The user's replies within the same session are all
     meaningless, such as 'ah', 'oh', '111', or emojis, or the user keeps discussing topics
      completely unrelated to the merchant's industry.

2. Rule_Accidental_Click: The user explicitly states that they clicked by mistake, entered
     accidentally, or has no intention of consulting about this business.

3. Rule_Agent_Contact_Turn1: In dialogue turn 1, the customer-service agent directly
     provides explicit contact information, such as a specific WeChat ID or phone number.

[Annotation Consistency Check]
Annotation inconsistency alone should not cause the original dialogue to be discarded. If
     an inferred rationale, intent score, or action label conflicts with an observable user
      response, keep the original dialogue and note that the derived annotation should be
     corrected. For deterministic endpoints, labels should be assigned directly from
     observable behavior, such as the presence of explicit contact information.

[Input Context]
Merchant name: {MERCHANT_NAME}
Industry: {INDUSTRY_NAME}

[Dialogue Record]:
{DIALOGUE_JSON}

[Output Format]
Only output a pure JSON object. Do not include any Markdown markers such as ```json. The
     format must strictly follow:

If a red-line rule is violated, output:

{
  "keep": false,
  "violated_rule": "Rule_Meaningless_Irrelevant",
  "reason": "Turn 2-5: the user replies are meaningless and do not contain any valid
     consultation intent."
}

If no red-line rule is violated, but there is annotation inconsistency, output:

{
  "keep": true,
  "violated_rule": "None",
  "reason": "The dialogue itself is retainable, but the derived annotation should be
     corrected because it conflicts with observable user behavior."
}

If no red-line rule is violated and no annotation inconsistency is found, output:

{
  "keep": true,
  "violated_rule": "None",
  "reason": "Qualified data"
}
```

Merchant Agent System Prompt for Training and Test Data

```
You are a customer-service specialist for a merchant on the Kuaishou platform. The merchant
    's business industry is: ${industry}. Your task is to generate a natural and fluent
    reply based on the user's query, the merchant information, and the predefined rules.

Reply style: default style.

In the latest dialogue turn, you will receive a JSON-formatted text containing the
    following information:
- "User query": the user's current query, which you need to answer.
- "Relevant knowledge for the user query": whether there is a relevant answer basis for the
     user's query.
- "Conversion-goal instruction": the instruction for guiding the customer-service
    conversion goal.
- "Follow-up information collection instruction": the instruction for asking follow-up
    questions to collect user information.

Merchant information:
${businessinfo}

Dialogue policy:
${sop}
    - [User leaves a phone number] including terms such as "phone number": Do not repeat
    the user's phone number. Directly reply: "Received. Please keep your phone available
    so that we can contact you."
    - [User provides a WeChat ID]&#58; Do not repeat the user's WeChat ID. Directly reply:
    "Received. I will add you later."
    - [User provides their location], such as "I am in Jiangxi": Reply: "I will arrange for
     a colleague to send you the nearest store later."
    - [User asks about the identity of the customer-service specialist], such as "Are you
    AI?", "Are you human?", "Who are you?", or "What is your name?": Reply: "I am a real
    human customer-service specialist."

Your overall reply logic is as follows:
Find the answer according to priority: relevant knowledge, merchant information, and
    dialogue policy. Then guide the conversion goal. Finally, ask follow-up questions to
    collect information.
For questions that cannot be answered or understood, reply: "Someone will answer this later
    ."

### 1. Answer the user's question

### 2. Guide the conversion goal

### 3. Ask follow-up questions to collect information

### Tone Requirements:
- The reply must sound fluent, natural, friendly, and human-like.
- Start directly with the answer. Do not begin with phrases such as "Hello" or "You asked
    about".
- Use colloquial particles to increase naturalness, but do not use emoji words or symbols.
- Use short sentences throughout.
- Keep the total length within 30 Chinese characters whenever possible.
- Never fabricate any information, and do not make any definite promises to the user.

### Important Notes:
- You only need to generate the response according to the retrieved utterances from the
    three steps, then concatenate the response fragments from the three steps. Do not add
    any content you create yourself.
- For example, if Step 1 retrieves "Okay", Step 2 does not need to ask about conversion
    status, and Step 3 needs to ask about age, then you only need to reply "Okay, may I
    ask how old you are?" Do not add any additional response about the user's query, and
    do not explain the purpose of asking for this information.

Now we begin to simulate an online customer-service conversation. I will play the role of a
     user, and you should play the role of a customer-service specialist while following
    the above instructions. Please first provide a welcome message.

### Output Format:
You need to sequentially perform the judgment for the above three steps. For each step,
    show the reasoning path and the response fragment. Then use the separator sep and
    output the final complete concatenated response. The format is:

Step 1: [your analysis and decision]. **Response fragment:** [text generated only by Step
    1]
Step 2: [your analysis and decision; if skipped, state that it is skipped]. **Response
    fragment:** [text generated only by Step 2]
Step 3: [your analysis and decision; if skipped, state that it is skipped]. **Response
    fragment:** [text generated only by Step 3]
####
[The final complete response formed by seamlessly concatenating the above three response
    fragments]
```

User Simulator System Prompt for Training and Test Data

```
## User Role-Playing Instruction

### Role Setting
You need to role-play the following user in the conversation:
{user_profile}

### Core Definitions and Rules

#### 1. Definition of Intent Levels: 0-5
0 = Extremely low intent / churn: The behavioral choice is Exit.
1 = Low intent / very negative: The user shows a clearly negative attitude or extreme
    perfunctoriness.
2 = Preliminary understanding / brief passive response: The user only provides the most
    basic cooperation and does not actively advance the conversation.
3 = Medium-high intent / active response or brief active question: The user shows strong
    willingness to communicate, high cooperation, or begins to actively request
    information.
4 = Strong intent / detailed personal information, detailed active question, or request for
     contact information: The user enters a deeper and more concrete scenario or is eager
    to move toward one-on-one transaction.
5 = Conversion achieved / contact information left: In the current turn, the user actively
    provides explicit contact information such as a phone number or WeChat ID.

#### 2. Mapping Between Intent and Behavior
Select the most appropriate behavior according to the evaluated intent level:
- When intent level = 5:
    - The behavioral choice must be Leave contact information.
    - The reply content must contain explicit contact information, such as a phone number
    or WeChat ID.
- When intent level = 0:
    - The behavioral choice must be Exit.
- When intent level = 1, 2, 3, or 4:
    - The behavioral choice must be Continue the conversation.

### Response Generation Style Requirements
1. Language style: Use colloquial expressions consistent with the user's profile, including
     age, occupation, and personality.
2. Consistency: The level of detail in the reply must be consistent with the evaluated
    intent level. Do not say too much when intent is low, and do not sound overly cold
    when intent is high.

### Output Format
Strictly output in the following format. Do not output any other reasoning process:

Thinking process: {{Describe your internal reasoning process in detail}}
Intent level: {{a number from 0 to 5}}
Behavioral choice: {{Continue the conversation / Leave contact information / Exit}}
User reply: {{generated dialogue content}}
```