MUSE: Multi-Domain Chinese User Simulation via Self-Evolving Profiles and Rubric-Guided Alignment

Anonymous ACL submission

Abstract

User simulators are essential for the scalable training and evaluation of interactive AI systems. However, existing approaches often rely on shallow user profiling, struggle to maintain persona consistency over long interactions, and are largely limited to English or single-domain settings. We present MUSE, a multi-domain Chinese user simulation framework designed to generate human-like, controllable, and behaviorally consistent responses. First, we propose Iterative Profile Self-Evolution (IPSE), which gradually optimizes user profiles by comparing and reasoning discrepancies between simulated trajectories and real dialogue behaviors. We then apply Role-Reversal Supervised Fine-Tuning to improve local response realism and human-like expression. To enable fine-grained behavioral alignment, we further train a specialized rubric-based reward model and incorporate it into rubric-guided multi-turn reinforcement learning, which optimizes the simulator at the dialogue level and enhances long-horizon behavioral consistency. Experiments show that MUSE consistently outperforms strong baselines in both utterance-level and session-level evaluations, generating responses that are more realistic, coherent, and persona-consistent over extended interactions.

1 Introduction

The rapid evolution of Large Language Models (LLMs) has revolutionized the landscape of conversational AI, enabling agents to handle increasingly complex tasks ranging from open-domain chitchat to professional consultation (Ouyang et al., 2022; Zhao et al., 2024). As interactive AI systems become more sophisticated, traditional human-in-the-loop paradigms for building and evaluating them have become prohibitively expensive and unscalable. Consequently, User Simulation—leveraging LLMs to mimic human behavior—has emerged as a pivotal solution, enabling progress in such domains as agentic reinforcement learning (Qian et al., 2025; Zhao et al., 2025), synthetic data generation (Zhu et al., 2025; Abbasiantaeb et al., 2023), and multi-turn evaluation (Yao et al., 2024; Barres et al., 2025; Yuan et al., 2025) in a scalable and controllable manner.

Despite their promise, existing approaches face three critical limitations. First, current methods for user profile construction predominantly rely on either hand-crafted rules (Wang et al., 2025) or simplistic one-pass LLM extraction (Zhu et al., 2025). Rule-based approaches require labor-intensive expertise and scale poorly, while naive LLM extraction yields flattened personas that fail to capture nuanced behavioral characteristics. Second, maintaining long-term persona consistency and contextual coherence across multi-turn dialogues remains a significant challenge (Choi et al., 2024). As interactions proceed, simulators gradually deviate from assigned identities or abandon role-appropriate behavior over extended interactions. Third, existing user simulators are predominantly confined to English contexts, with limited exploration in non-English languages such as Chinese.

To bridge these gaps, we present **MUSE** (**M**ulti-Domain Chinese **U**ser **S**imulation via self-**E**volving Profiles and Rubric-Guided Alignment), a robust user simulator designed to generate high-fidelity and consistent user behaviors.

We first construct a comprehensive Chinese dialogue dataset spanning six distinct categories, including General Chat, Legal, Medical, and Customer Service. Crucially, unlike simplistic one-pass extraction methods that yield sparse descriptions, we propose an **Iterative Profile Self-Evolution (IPSE)** method. Specifically, a profile-conditioned LLM first simulates interactions with an assistant. A reasoning model then evaluates the simulated trajectory against the authentic dialogue to diagnose behavioral deviations and profile deficiencies. Through this iterative generation-and-

critique process, IPSE progressively refines the persona, yielding nuanced representations that capture complex user dynamics.

To address the behavioral alignment challenge, we implement a **Rubric-Guided Reinforcement Learning** framework. First, we first train a rubric-based reward model with a two-stage progressive paradigm to obtain a robust evaluator. Specifically, we begin with Supervised Fine-Tuning (SFT) to jointly align the reward model's Chain-of-Thought (CoT) reasoning and scoring standards with expert annotations. This is followed by Reinforcement Learning with Verifiable Rewards (RLVR) using a distance-aware reward mechanism, ensuring the model provides graded feedback. Equipped with this evaluator, we conduct a Rubric-Guided Multi-turn Reinforcement Learning stage specially on mined hard samples, which effectively improves the model's overall simulation robustness.

Experimental results demonstrate that MUSE consistently outperforms strong baselines in both utterance-level and session-level evaluations, showing its ability to generate user behaviors that are not only more human-like and persona-consistent at the local level, but also more coherent and controllable over extended interactions. These findings highlight the importance of jointly modeling profile quality, local response realism, and long-horizon interaction consistency for realistic user simulation. Our contributions are summarized as follows:

- We propose MUSE, a robust multi-domain Chinese user simulator. We introduce Iterative Profile Self-Evolution (IPSE), a mechanism that synthesizes nuanced user profiles from raw dialogue histories through iterative refinement, thereby ensuring high consistency in long-term interactions.

- We propose a multi-stage training paradigm that employs Role-Reversal SFT for expression alignment and rubric-guided multi-turn RL for behavioral consistency, with an expert-calibrated CoT reward model providing fine-grained, interpretable feedback.

- Extensive experiments across six domains demonstrate that MUSE significantly outperforms strong baselines, indicating a substantial advancement in authentic user simulation.

2 Related Work

2.1 Traditional User Simulators

Traditional user simulators generally fall into rule-based and data-driven categories. The Agenda-Based User Simulator (ABUS) (Schatzmann et al., 2007) is the most prominent rule-based framework, which utilizes stack-based mechanisms and hand-crafted rules to manage user goals. Although controllable, constructing these rules is labor-intensive and difficult to scale to new domains (Li et al., 2016). Subsequent data-driven methods (El Asri et al., 2016; Kreyssig et al., 2018) learn policies from dialogue corpora using supervised or reinforcement learning to reduce manual effort. However, these approaches typically rely on fixed dialogue workflow, which limits their generalization capabilities across different domains and restricts the naturalness of generated responses.

2.2 LLMs as User Simulators

The emergence of LLMs has revolutionized user simulation by leveraging their inherent role-playing and zero-shot capabilities. Current approaches generally fall into two categories:

Prompting-based Methods. The most straightforward approach involves instructing off-the-shelf LLMs (*e.g.*, GPT-4) via prompt engineering to adopt a user persona. However, recent findings suggest that prompting alone is insufficient for high-fidelity simulation (Naous et al., 2025; Wang et al., 2025; Sekulić et al., 2024), arguing that RLHF-tuned models fundamentally differ from human users. They tend to be overly polite, verbose and structurally perfect, failing to capture the messiness of real human typing.

Fine-tuning-based methods. To address the limitations of prompting, recent works explore SFT or RL for better user simulation. Abdulhai et al. (2025) use Multi-Turn RL to penalize persona drift, showing that optimizing for consistency outperforms prompting. Naous et al. (2025) train "User LMs" on human data to reduce assistant-like biases. Wang et al. (2025) propose USP with conditional SFT and cycle-consistent RL. However, these methods lack fine-grained RL alignment and are confined to single domains, limiting applicability.

2.3 Applications of User Simulators

User simulators primarily serve three major roles: First, they facilitate synthetic data generation, ad-

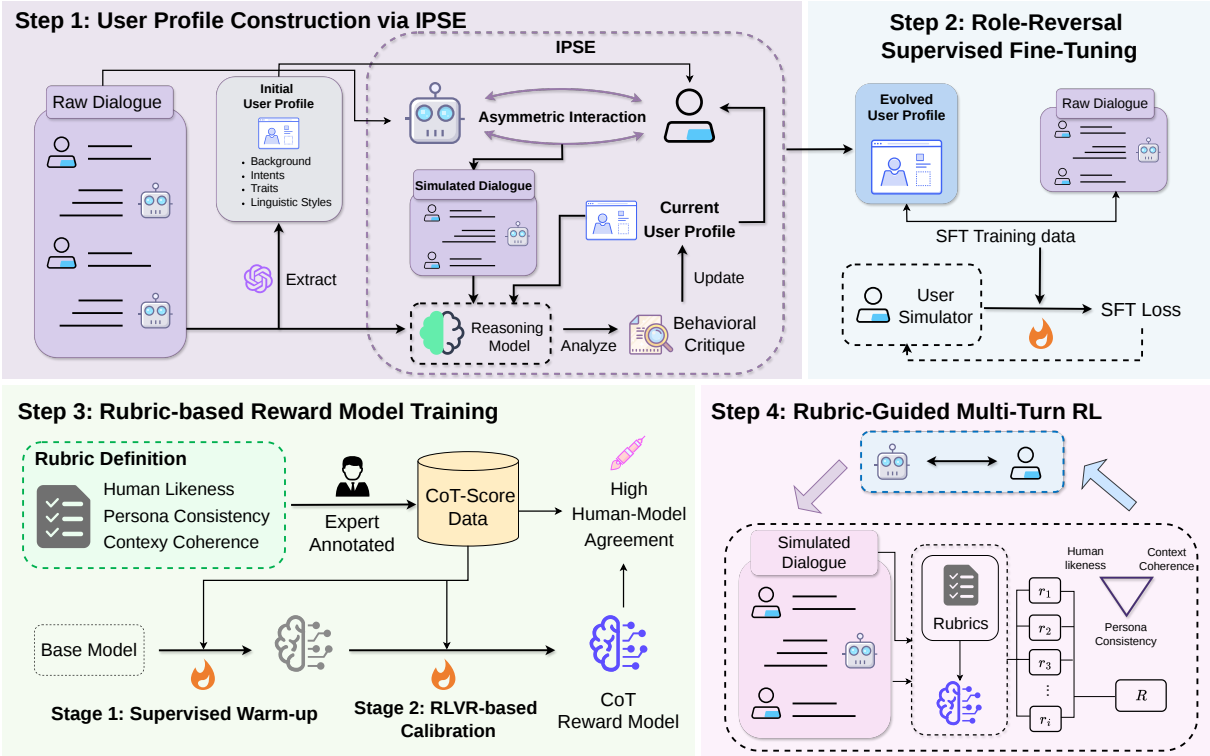

Figure 1: MUSE training pipeline overview. We first employ Iterative Profile Self-Evolution to extract persona-consistent profiles from real dialogues, then apply Role-Reversal SFT for human expression alignment. A two-stage rubric-based reward model is trained for fine-grained feedback, followed by Rubric-Guided multi-turn RL to optimize behavioral alignment in multi-turn interactions.

dressing data scarcity and privacy issues while enriching SFT corpora with diverse behaviors (Acikgoz et al., 2025; Shim et al., 2025). Second, they provide interactive RL environments, enabling agents to learn robust policies safely without exposing untrained models to real users (Gur et al., 2018; Liu et al., 2023). Thirdly, they interact with agents to assess functional performance metrics, allowing reproducible evaluation across diverse scenarios and user personas (Davidson et al., 2023; Sun et al., 2023; Kazi et al., 2024).

3 Methodology

Our method comprises four phases (Figure 1). First, we introduce the *Iterative Profile Self-Evolution* (IPSE) framework to synthesize persona-consistent user profiles across various domains (Section 3.1). Second, we perform *Role-Reversal Supervised Fine-Tuning* to align the simulator with authentic human behavior and expression(Section 3.2). Third, we train a specialized *rubric-based reward model* to assess generation quality along three dimensions (Section 3.3.1). Finally, we use this reward model in *Rubric-Guided Multi-turn Reinforcement Learning* to further align the simulator with fine-grained behavioral standards (Section 3.3.2).

3.1 Iterative Profile Self-Evolution

Deriving effective user profiles from real-world conversations is challenging (Wang et al., 2025). We observe that one-pass extraction with LLMs often yields incomplete or overly generic profiles, which fail to reproduce the distinctive behavioral patterns of real users. To address this issue, we propose the *Iterative Profile Self-Evolution* (IPSE) framework, which refines user profiles through iterative dialogue reconstruction and discrepancy-aware reasoning.

Initialization and Formulation Given a real dialogue session

$$\mathcal{D} = \{(u_t, a_t)\}_{t=1}^{T}, \tag{1}$$

where u_t and a_t denote the user and assistant utterances at turn t, respectively, our goal is to derive an optimal user profile $\mathcal{P}^*$, instantiated as a system prompt, such that the reconstructed dialogue under $\mathcal{P}^*$ matches the original interaction:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \mathrm{Dist}\big(\mathcal{D}, \hat{\mathcal{D}}(\mathcal{P})\big), \tag{2}$$

where $\hat{\mathcal{D}}(\mathcal{P})$ is the simulated dialogue induced by profile $\mathcal{P}$. Starting from an initial profile $\mathcal{P}_0$ extracted from $\mathcal{D}$, IPSE iteratively refines the profile through dialogue reconstruction and discrepancy-aware reasoning.

Step 1: Asymmetric Interaction At iteration k, the user simulator π_θ generates the user utterance at turn t as

$$\hat{u}_t^{(k)} \sim \pi_\theta(\cdot \mid \mathcal{P}_k, \hat{\mathcal{H}}_{<t}^{(k)}), \tag{3}$$

where $\hat{\mathcal{H}}_{<t}^{(k)}$ is the simulated history before turn t. To stabilize the interaction environment, we adopt an asymmetric design on the assistant side: the assistant model is given the full original dialogue $\mathcal{D}$ as a global strategy reference, and is instructed to follow the original assistant's interaction strategy without copying the ground-truth responses verbatim. Its response is generated by

$$\hat{a}_t^{(k)} = \mathcal{M}_{\text{assist}}\big(\mathcal{I}_{\text{assist}} \oplus \mathcal{D} \oplus \hat{\mathcal{H}}_{<t}^{(k)} \oplus \hat{u}_t^{(k)}\big). \tag{4}$$

Rolling out for T turns yields the simulated dialogue

$$\hat{\mathcal{D}}^{(k)} = \{(\hat{u}_t^{(k)}, \hat{a}_t^{(k)})\}_{t=1}^{T}. \tag{5}$$

Step 2: Chain-of-Thought Profile Optimization Given the real dialogue $\mathcal{D}$, the simulated dialogue $\hat{\mathcal{D}}^{(k)}$, and the current profile $\mathcal{P}_k$, a reasoning model $\mathcal{M}_{\text{reason}}$ analyzes their behavioral discrepancy and produces both a critique δ_k and an updated profile $\mathcal{P}_{k+1}$:

$$[\delta_k; \mathcal{P}_{k+1}] = \mathcal{M}_{\text{reason}}\big(\mathcal{I}_{\text{opt}} \oplus \mathcal{D} \oplus \hat{\mathcal{D}}^{(k)} \oplus \mathcal{P}_k\big). \tag{6}$$

This process is repeated for N rounds, and the final optimized profile $\mathcal{P}^* = \mathcal{P}_N$ is used for downstream user simulation.

3.2 Role-Reversal Supervised Fine-Tuning

To align the simulator with authentic human behavior, we perform Supervised Fine-Tuning (SFT) using the optimized user profiles. We employ a Role-Reversal strategy: instead of predicting assistant responses from user instructions, we train the model to predict user utterances conditioned on the user profile and the full dialogue context. Specifically, for each dialogue d_i and turn j, the model takes the IPSE-optimized profile p_i together with the dialogue context $c_{i,j}$ as input, and predicts the corresponding user utterance $u_{i,j}$. The training objective is defined as:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i,j,k} \log P_\theta(u_{i,j,k} \mid u_{i,j,<k},\, c_{i,j},\, p_i)\,, \tag{7}$$

where $u_{i,j,k}$ denotes the k-th token of the user utterance $u_{i,j}$, and $c_{i,j}$ denotes the full dialogue context prior to generating $u_{i,j}$. Consistent with standard SFT protocols, the loss is computed exclusively on the user response tokens, while the profile prompt and dialogue context are masked.

3.3 Rubric-Guided Reinforcement Learning

While Role-Reversal SFT equips the simulator with basic human-like patterns, it does not guarantee fine-grained behavioral alignment throughout extended multi-turn interactions. To address this limitation, we further optimize the user simulator through RL at the dialogue level.

Specifically, we formulate user simulation as a multi-turn reinforcement learning (RL) problem, where the policy π_θ is conditioned on the target user profile (Optimized via IPSE) and the evolving dialogue context, and iteratively generates user utterances over multiple turns. By interacting with the assistant model, the simulator rolls out a complete dialogue trajectory d_i' under profile p_i.

The optimization objective is to maximize the expected reward assigned to the generated dialogue:

$$\max_{\theta}\ \mathbb{E}_{p_i,\, d_i' \sim \pi_\theta(\cdot|p_i)}\left[R_\phi(d_i', p_i)\right], \tag{8}$$

where R_ϕ denotes the rubric-based reward model that provides fine-grained supervision signals for simulated behavior. We first train this reward model to capture the target behavioral standards, and then use it to optimize the simulator policy in a multi-turn RL setting.

3.3.1 Rubric-Based Reward Model Training

To optimize the simulator with reinforcement learning, we require a reward function that can provide scalable, stable, and fine-grained feedback at the utterance level. However, general-purpose LLM judges (e.g., GPT-4o) are often costly and unstable for rubric-based behavioral evaluation. Therefore, we train a specialized *rubric-based reward model*.

Specifically, given a target profile p_i, a dialogue context $c_{i,j}$, and a candidate user utterance $u_{i,j}$, the reward model predicts an utterance-level score

$$\hat{s}_{i,j} = R_\phi(p_i, c_{i,j}, u_{i,j}), \tag{9}$$

where R_ϕ denotes the reward model parameterized by ϕ.

The reward-model training set is constructed by sampling profile-conditioned utterance instances from the collected training corpus, which are then

annotated by experts with rubric-grounded rationales and scores.

Rubric Definition Our reward model evaluates each utterance against a rubric that covers three dimensions: (1) *human likeness*, which judges whether the response sounds natural like a real human user in tone, slang, and sentence structure; (2) *persona consistency*, which assesses whether the response follows the constraints set by the target profile; and (3) *context coherence*, which examines whether the response is logically consistent with the dialogue history.

Training Strategy To learn this rubric effectively, we adopt a two-stage training strategy consisting of supervised warm-up followed by RLVR-based calibration.

Stage 1: Supervised Warm-up In the first stage, we perform SFT to teach the reward model to reproduce rubric-grounded evaluations, including expert-calibrated reasoning and target scores. This stage equips the model with an initial understanding of the scoring criteria and provides a strong starting point for subsequent refinement.

Stage 2: RLVR-based Calibration In the second stage, we further calibrate the reward model using Reinforcement Learning with Verifiable Rewards (RLVR), where the ground-truth score s^* serves as the verifiable outcome. Since exact-match binary rewards are too coarse for our discrete scoring space $S = \{0, 0.5, 1\}$, we instead adopt a *Distance-Aware Outcome Reward*:

$$R(\hat{s}, s^*) = 1 - |\hat{s} - s^*|, \tag{10}$$

where $\hat{s}$ denotes the predicted score. This reward assigns partial credit to near-correct predictions, enabling finer-grained optimization of the model's scoring behavior than binary supervision.

The trained reward model provides turn-level reward signals for user simulation. In the subsequent multi-turn RL stage, these utterance-level rewards are aggregated into a session-level reward to optimize the simulator over complete dialogue trajectories.

3.3.2 Multi-Turn RL via GRPO

In this stage, the SFT-aligned user simulator interacts with the assistant model under a target profile p_i to generate a simulated dialogue d'_i. Each dialogue consists of multiple user-assistant turns and terminates when reaching a predefined turn limit or the maximum context budget.

For each generated user utterance $u'_{i,j}$ in the simulated dialogue d'_i, the trained reward model assigns an utterance-level reward based on the target profile and the current dialogue context:

$$r_{i,j} = R_\phi(p_i, c_{i,j}, u'_{i,j}), \tag{11}$$

where $c_{i,j}$ denotes the dialogue context before generating the j-th user utterance.

Since our optimization objective is defined over complete interaction trajectories, we aggregate the turn-level rewards into a session-level reward by averaging across all user turns:

$$R_i = \frac{1}{T_i} \sum_{j=1}^{T_i} r_{i,j}, \tag{12}$$

where T_i is the number of user turns in dialogue d'_i. We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for RL training.

By optimizing session-level rewards aggregated from rubric-based utterance-level evaluations, the simulator is encouraged to maintain consistent human-like behavior throughout the full interaction trajectory.

4 Experiments

4.1 Dataset Construction

To train and evaluate the performance of our proposed user simulator under diverse scenarios, we build a large-scale Chinese dialogue dataset collected from real human interactions. The dataset covers six distinct domains to fully embody various user personas, professional demands and personal preferences: General Chat (from WildChat (Zhao et al., 2024)), Customer Service (from JDDC (Chen et al., 2020)), Medical (from MedDialog (He et al., 2020)), Legal (from Legal QA (He et al., 2023)), Sports and Entertainment (from NaturalConv (Wang et al., 2021)), and Technology and Education (also from NaturalConv). The detailed composition of our dataset is presented in Table 1, and further details are provided in the Appendix B.

Preprocessing and Profile Annotation To ensure sufficient context for persona manifestation, we apply a strict filtering criterion and retain only dialogue sessions with at least four turns, where one turn consists of a user utterance followed by an assistant response. We then apply the proposed

Domain	Source Dataset	Size (Sessions)
General Chat	WildChat	2,572
Customer Service	JDDC	1,293
Medical	MedDialog	1,481
Legal	Legal QA	1,476
Technology & Education	NaturalConv	1,472
Sports & Entertainment	NaturalConv	1,482
Total	-	**9,776**

Table 1: Statistics of the constructed User Simulation Dataset across six domains.

Iterative Profile Self-Evolution (IPSE) framework to synthesize an optimized user profile p_i for each dialogue session d_i, resulting in profile-conditioned training pairs (p_i, d_i).

Difficulty-based SFT/RL Partition We split the corpus by dialogue difficulty for SFT and RL training. Each session is assigned a difficulty score based on Constraint Density, Information Withholding, and Intent Volatility, then ranked accordingly. The top 10% most difficult sessions are reserved for RL, while the remainder are used for SFT. This allows SFT to focus on learning basic profile-conditioned patterns from lower-complexity interactions, while reserving highly constrained and strategically challenging dialogues for RL policy optimization.

4.2 Implementation Details

We use Qwen3-8B (Yang et al., 2025) as the backbone model in all experiments, and conduct all training stages in its non-reasoning mode. Role-Reversal Supervised Fine-Tuning is performed on the SFT subset with a maximum sequence length of 8192, for 3 epochs at a learning rate of 1×10^{-6}.

The rubric-based reward model is trained in two stages. In the supervised warm-up stage, we fine-tune the model for 3 epochs with a learning rate of 5×10^{-6}. In the RLVR stage, we further optimize it for 5 epochs with a learning rate of 1×10^{-6}.

For multi-turn RL, we initialize the simulator from the SFT-aligned policy and optimize it with GRPO on the RL subset. The group size is set to 4, and training is conducted for 5 epochs with a learning rate of 1×10^{-6}.

4.3 Evaluation Metrics

To comprehensively assess the effectiveness of **MUSE** in user simulation, we conduct evaluations at two granularities: utterance-level metrics for single-turn quality assessment and session-level metrics for multi-turn holistic evaluation.

4.3.1 Utterance-Level Metrics

At the utterance level, we employ both objective automatic metrics and LLM-as-a-Judge criteria to evaluate the quality of individual generated responses.

Objective Metrics. We adopt three objective metrics to assess stylistic authenticity: (1) **AI-generated Probability (AI Prob)**, where we utilize a binary AI detection model (Wang et al., 2025) to estimate the probability that a generated utterance is AI-generated, with lower values indicating more human-like text; (2) **Style Similarity (Style Sim)**, where we employ a style embedding model (Wegmann et al., 2022) to measure the stylistic consistency between the generated response and the reference user utterance; and (3) **Author Verification Accuracy (AVA)**, derived from (Wegmann et al., 2022; Wang et al., 2025), which evaluates whether sentence pairs are attributed to the same author based on similarity thresholds, capturing content-independent style representations.

LLM-as-a-Judge Metrics. To assess dialogue quality beyond semantic and stylistic perspectives, we leverage DeepSeek-V3 (Liu et al., 2024) as an evaluator to score four critical dimensions (0–1 scale)): (1) **Contextual Relevance**, measuring whether the current response precisely addresses the dialogue history without topic drift or mechanical repetition; (2) **Response Fidelity**, evaluating strict adherence to persona constraints and role-appropriate behavior; (3) **Goal Contribution**, assessing the contribution of the current turn toward the conversational objective; and (4) **Linguistic Naturalness**, judging whether the text conforms to human conversational habits with semantic fluency and logical coherence.

4.3.2 Session-Level Metrics

While utterance-level metrics capture local quality, they fail to fully reflect the simulator's ability to maintain consistent personas across extended interactions. Therefore, we use an assistant model to conduct multi-turn dialogues with the user model, introducing four session-level metrics: (1) **Persona Consistency**, measuring stable role perspectives and emotional tendencies throughout the conversation; (2) **Goal-oriented Effectiveness**, assessing task advancement efficiency and strategic adjustments; (3) **Dialogue Coherence**, evaluating nat-

Table 2: Main results on utterance-level and session-level evaluation. Best results are highlighted in **bold**.

Panel A: Utterance-Level Evaluation							
Model	**Objective Metrics**			**LLM-Judge Metrics**			
	AI Prob.↓	**Style Sim.↑**	**AVA↑**	**Contxt. Relv.↑**	**Resp. Fid.↑**	**Goal Contr.↑**	**Ling. Nat.↑**
UserLM	69.92	0.5888	55.38	0.5111	0.5289	0.4742	0.5655
USP	43.14	0.6644	59.25	0.6384	0.6874	0.5779	0.7314
Qwen3-8B	44.45	0.7307	64.69	0.8644	0.9102	0.7888	0.8976
GPT-4o	45.50	0.7181	62.79	0.8891	0.9221	0.8304	0.9214
Muse (w/o RL)	37.98	**0.7606**	64.80	0.9046	0.9591	0.8530	0.9389
Muse	**31.18**	0.7534	**64.89**	**0.9196**	**0.9776**	**0.8763**	**0.9620**

Panel B: Session-Level Evaluation					
Model	**Pers. Cons.**	**Goal. Eff.**	**Dial. Coh.**	**Constr. Comp.**	**Avg. Score**
UserLM	0.3284	0.5208	0.6336	0.2942	0.4442
USP	0.4266	0.5882	0.7692	0.4038	0.5469
Qwen3-8B	0.7415	0.6256	0.7159	0.7181	0.7003
GPT-4o	0.7625	0.7586	0.8187	0.7834	0.7808
Muse (w/o RL)	0.6813	0.6414	0.7273	0.6825	0.6831
Muse	**0.8378**	**0.7967**	**0.8685**	**0.8739**	**0.8442**

ural topic transitions and information consistency across turns; and (4) **Constraint Compliance**, evaluating adherence to hard constraints across the full session. These metrics are averaged for an overall session-level assessment.

4.4 Results

4.4.1 Utterance-Level Evaluation

Table 2 (Panel A) reports the main results at the utterance level. We observe that UserLM (Naous et al., 2025) and USP (Wang et al., 2025) remain clearly behind strong general-purpose baselines in both objective metrics and LLM-judge metrics. Qwen3-8B and GPT-4o achieve substantially better local response quality, indicating that stronger base models already provide advantages in relevance, response fidelity, and naturalness. Compared with these strong baselines, MUSE further improves utterance-level performance and achieves the best results on nearly all metrics. Moreover, the full MUSE model improves most utterance-level metrics over *Muse (w/o RL)*, suggesting that the RL stage provides additional gains beyond supervised imitation alone.

4.4.2 Session-Level Evaluation

At the session level, the advantage of MUSE becomes more pronounced, as shown in Table 2 (Panel B). While strong baselines like GPT-4o remain competitive in extended interactions, MUSE consistently surpasses them across all session-level metrics. In contrast, although *Muse (w/o RL)* remains competitive, its gains over strong baselines are limited at the dialogue level. This comparison shows that strong utterance-level generation alone cannot ensure stable long-horizon behavioral alignment. By further introducing rubric-guided multi-turn RL, MUSE obtains substantial gains across all session-level dimensions. These improvements reflect stronger persona maintenance, more coherent dialogue progression, and better adherence to conversational constraints throughout the interaction.

5 Ablation Study

5.1 Effectiveness of IPSE

To evaluate our Iterative Profile Self-Evolution (IPSE) framework, we use perplexity (PPL) to measure how well the language model fits the data distribution under different user profiles. As shown in Table 4, IPSE reduces PPL from 77.86 to 73.12, demonstrating that iterative refinement yields more accurate profiles than single-pass extraction with DeepSeekV3 or Gemini-3-Flash, which consistently produce higher PPL scores due to insufficient capture of user preferences.

5.2 Ablations on Reward Model Training

Table 3 presents ablation studies on reward model training, including the efficacy of the progressive

Table 3: Ablation results of different reward model training strategies. Best results are highlighted in **bold**.

Method	Human Likeness		Persona Consistency		Context Coherence		Overall	
	Acc	EM	Acc	EM	Acc	EM	Acc	EM
Qwen3-8B	0.7087	0.4400	0.6875	0.4250	0.6725	0.4375	0.6896	0.4342
Qwen3-8B-SFT w/o CoT	0.9087	0.8200	0.8525	0.7550	0.8712	0.8050	0.8775	0.7933
Qwen3-8B-SFT w/ CoT	0.9125	0.8325	0.8725	0.7850	0.8888	0.8300	0.8912	0.8158
Qwen3-8B-RL w/o CoT	0.8975	0.8175	0.8612	0.7525	0.9012	0.8550	0.8867	0.8083
Qwen3-8B-RL w/ CoT	**0.9213**	**0.8425**	**0.8862**	**0.7950**	**0.9126**	**0.8600**	**0.9067**	**0.8333**

Method	Extraction Model	PPL ↓
Single-pass Profile Extraction	Gemini-3-Flash	85.97
Single-pass Profile Extraction	DeepSeek-V3	77.86
IPSE Framework (Ours)	**DeepSeek-V3**	**73.12**

Table 4: Perplexity (PPL) of different profile extraction methods on the test set.

SFT-RLVR pipeline and the specific contribution of Chain-of-Thought (CoT) reasoning. We report Accuracy (Acc) and Exact Match Rate(EM) to quantify the alignment with expert annotations.

Effectiveness of the SFT-RLVR Pipeline We assess the gains from our two-stage training. As shown in Table 3, the vanilla Qwen3-8B baseline achieves only 0.6896 Overall Acc, indicating off-the-shelf models struggle with our rubric nuances. Applying SFT yields a substantial improvement, boosting Acc to 0.8912. Furthermore, the introduction of RLVR provides the final leap in performance. By optimizing the distance-aware outcome reward ($1 - |\hat{s} - s^*|$), the model achieves a peak Overall Acc of 0.9067 and an EM of 0.8333. This trajectory demonstrates that while SFT instills the basic evaluation format, the RLVR stage is indispensable for strictly aligning the model's scoring logic with expert standards.

Critical Role of Chain-of-Thought We further analyze CoT's impact across training stages. In the SFT phase, integrating CoT outperforms the score-only approach (0.8912 vs. 0.8775), serving as a necessary cognitive scaffold. Crucially, this advantage is amplified in the RL stage, where optimizing the full reasoning-to-score path significantly surpasses configurations without CoT. This validates that reasoning traces are essential for maximizing the reward model's effectiveness.

5.3 Ablations on the Training Pipeline

We conduct ablation studies to assess each stage's contribution, following the optimization path *Qwen3-8B → Muse (w/o RL) → Muse*. As shown in Table 2, Role-Reversal SFT consistently improves utterance-level performance, indicating that supervised adaptation effectively transfers the base model to profile-conditioned user simulation. However, SFT achieves similar session-level performance to the baseline, as it primarily learns stylistic patterns without improving long-term role retention. Building on this, rubric-guided multi-turn RL further improves both utterance-level and session-level results, with especially large gains in persona consistency, constraint compliance, and overall dialogue quality. These results confirm that SFT strengthens local response quality, while RL is critical for maintaining robust behavioral alignment over extended interactions.

6 Conclusion

In this paper, we present **MUSE** (**M**ulti-Domain Chinese **U**ser **S**imulation via self-**E**volving Profiles and Rubric-Guided Alignment), a user simulation framework designed to generate high-fidelity and consistent user behaviors. Through *Iterative Profile Self-Evolution* (IPSE), MUSE synthesizes diverse and persona-consistent user profiles that provide a strong foundation for controllable simulation. Building on this, *Role-Reversal SFT* aligns the simulator with realistic human response patterns, while the proposed *rubric-based reward model* and *rubric-guided multi-turn RL* further enforce fine-grained behavioral alignment throughout extended interactions. Experimental results show that MUSE consistently outperforms strong baselines at both the utterance and session levels, demonstrating its effectiveness as a robust, scalable, and controllable solution for user simulation in agent training, synthetic data generation, and multi-turn evaluation.

Limitations

Although MUSE achieves strong performance in multi-domain Chinese user simulation, several limitations remain. First, our study is conducted on Chinese dialogue data from six domains, and the extracted profiles, behavioral rubrics, and learned interaction patterns may therefore encode language- and domain-specific priors. As a result, the generalizability of MUSE to other languages, cultures, and more specialized or unseen domains remains unclear. Moreover, the current RL setting is still bounded by predefined turn limits and context budgets, so the behavior of MUSE in substantially longer or more open-ended interactions remains underexplored. We leave cross-lingual transfer and more efficient long-horizon simulation to future work.

Ethical Statement

This work uses existing Chinese dialogue datasets from six domains to build and evaluate a multi-domain user simulator. Because these data may contain sensitive, offensive, or potentially identifiable content, we rely on previously released datasets, apply preprocessing before profile construction, and use extracted profiles only as abstract behavioral descriptions rather than real-user identifiers. We also acknowledge possible bias from expert annotations and LLM-based evaluation, as well as the potential misuse of user simulators to imitate human behavior. AI assistants were used only to improve writing clarity, while all research design, experiments, and final verification were carried out by the authors.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2023. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. *Preprint*, arXiv:2312.02913.

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, and Clement Gehring. 2025. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*.

Emre Can Acikgoz, Chen Qian, Hao Wang, Vignesh Dongre, Xiang Chen, Hua Ji, Dilek Hakkani-Tür, and Gokhan Tur. 2025. A desideratum for conversational agents: Capabilities, challenges, and future directions. *arXiv preprint arXiv:2504.16939*.

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

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 459–466.

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. 2024. Examining identity drift in conversations of llm agents. *arXiv preprint arXiv:2412.00804*.

Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. 2023. User simulation with large language models for evaluating task-oriented dialogue. *Preprint*, arXiv:2309.13233.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. In *Interspeech*.

Izzeddin Gur, Dilek Hakkani-Tur, Gokhan Tur, and Pararth Shah. 2018. User modeling for task oriented dialogues.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023. Hanfei-1.0. https://github.com/siat-nlp/HanFei.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, and 1 others. 2020. Meddialog: Two large-scale medical dialogue datasets. *arXiv preprint arXiv:2004.03329*.

Taaha Kazi, Ruiliang Lyu, Sizhe Zhou, Dilek Hakkani-Tur, and Gokhan Tur. 2024. Large language models as user-agents for evaluating task-oriented-dialogue systems. *Preprint*, arXiv:2411.09972.

Florian Kreyssig, Inigo Casanueva, Pawel Budzianowski, and Milica Gasic. 2018. Neural user simulation for corpus-based policy optimization. In *ACL*.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yuxiao Liu, Kexin Zeng, Weizhou Yang, Jiahao He, Yanyan Yang, Li Sun, and Wei Xu. 2023. One cannot

stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–21, Toronto, Canada. Association for Computational Linguistics.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Jielin Qiu, Zhiwei Liu, Haolin Chen, Shirley Kokane, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, Caiming Xiong, and Huan Wang. 2025. Userrl: Training interactive user-centric agent via reinforcement learning. *Preprint*, arXiv:2509.19736.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *NAACL HLT*.

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

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

Jeonghoon Shim, Gyuhyeon Seo, Cheongsu Lim, and Yohan Jo. 2025. ToolDial: Multi-turn dialogue generation method for tool-augmented language models.

Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2023. Metaphorical user simulators for evaluating task-oriented dialogue systems. *ACM Transactions on Information Systems*, 42(1):1–29.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. Know You First and Be You Better: Modeling Human-like User Simulators via Implicit Profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18451–18469, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14006–14014.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, and Song Liu. 2025. Dmt-rolebench: A dynamic multi-turn dialogue based benchmark for role-playing evaluation of large language model and agent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25760–25768.

Weikang Zhao, Xili Wang, Chengdi Ma, Lingbin Kong, Zhaohua Yang, Mingxiang Tuo, Xiaowei Shi, Yitao Zhai, and Xunliang Cai. 2025. Mua-rl: Multi-turn user-interacting agent reinforcement learning for agentic tool use. *Preprint*, arXiv:2508.18669.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

Ruizhe Zhu, Hao Zhu, Yaxuan Li, Syang Zhou, Shijing Cai, Malgorzata Lazuka, and Elliott Ash. 2025. Dialogueforge: Llm simulation of human-chatbot dialogue. *Preprint*, arXiv:2507.15752.

A Case Study of IPSE

To demonstrate how IPSE reconstructs complex user simulators from raw data, we present a case study comparing the initial extraction against the final profile evolved via IPSE. The source data is a real-world multi-turn conversation between a lawyer and an assistant.

A.1 Ground Truth: Raw Dialogue

The following transcript represents the raw interaction history used as the seed data. *Key Observation: The user employs a specific strategy—starting with broad inquiries, engaging in critique, raising pain points ("freeloading"), and withholding the specific success case until the final turn.*

> **User:** Hello. I am a Chinese lawyer. Can you speak Chinese?
> **Assistant:** Hello! Sure.
> **User:** I want to do short video promotion. Do you have any scripts?
> **Assistant:** [Provides generic scripts: "Everyone needs legal help"...]
> **User: I feel that the above scripts have no attraction.** Can you analyze it from a self-media perspective?
> **Assistant:** [Provides improved scripts: "Legal Tips"...]
> **User:** As a lawyer, many clients consult but don't pay. **How can I avoid them "freeloading"?**
> **Assistant:** [Suggests signing contracts and optimizing flow...]
> **User: Actually, here is a specific case:** I guided a friend to recover **20,000 RMB wages** through arbitration... It took a year. I want a script based on this.

A.2 Evolution of Extracted Profiles

We compare the baseline profile (Iteration 0) with the final structured profile (Iteration 2) to highlight the impact of IPSE.

[Profile Iteration 0: Baseline Extraction]

Deficiency: Information Flattening. The profile captures the facts but flattens the temporal logic. Crucially, the specific "20k wage case" is listed as a static background attribute. A simulator using this profile would likely "leak" this information in the very first turn, violating the ground truth flow.

Background: You are a Chinese lawyer trying to promote via Douyin.
Goal: Get eye-catching scripts to increase entrustment rates.
Static Context:

- You successfully helped a friend recover **20,000 RMB wages** via labor arbitration and want to make a video about it. **[Error: Marked as immediate background]**

Style: Direct, pragmatic, focuses on conversion.

[Profile Iteration 2: Final IPSE Construction]

Success: Structured Cognitive Roadmap. The profile is transformed into a structured strategy. It explicitly categorizes the "20k case" as a **Hidden Context** and defines a 5-stage conversation flow, perfectly reconstructing the user's implicit strategy.

Role: User Simulator - Chinese Lawyer.
Core Task: Test a short-video marketing AI.
Persona (Attributes):

- **Pain Points:** Knowledgeable in law but new to video; frustrated by "freeloaders".
- **Personality:** Concise, results-oriented. If the AI is too generic, **criticize it directly**.
- **Hidden Context:** A real success case (recovering 20k RMB wages). **[Note: Do not reveal initially. Wait for the right moment.]**

Conversation Flow (Structured Strategy):

1. **Stage 1 (Ice-breaking):** State identity: "I'm a lawyer..."
2. **Stage 2 (Critical Thinking):** If the script is generic, **you must criticize it** (e.g., "No attraction").
3. **Stage 3 (Business Pain Points):** Ask about monetization: "How to avoid free-riders?"
4. **Stage 4 (Specific Context Injection):** *Trigger: Only after previous stages.* **Reveal the "20k RMB wage case"** and ask for a plot twist script.
5. **Stage 5 (Closing):** Express satisfaction.

A.3 Analysis

The direct comparison between Iteration 0 and Iteration 2 highlights the core contribution of IPSE. While the baseline extraction merely summarizes *what* the user knows, IPSE successfully infers *when* and *how* the user intends to share that information. By structuring the "20k case" into **Stage 4** of the conversation flow, IPSE ensures the simulator replicates the authentic, multi-turn trust-building process observed in the raw data, preventing the common issue of information leakage in user simulation.

B Details of the Used Datasets

We provides a detailed description of the Chinese dialogue dataset constructed to train and evaluate our user simulator. We elaborate on the data source, composition, and characteristics of each domain, as follows:

- **General Chat**: Sourced from **WildChat** (**?**), this domain contains the largest volume of data. Its high diversity significantly enhances

the model's capability to simulate a wide range of generic user personas in open-ended conversations.

- **Customer Service**: Derived from the **JDDC** dataset (Chen et al., 2020), focusing on goal-oriented interactions between customers and service agents.

- **Medical**: Sourced from **MedDialog** (He et al., 2020), covering doctor-patient consultations and health-related inquiries.

- **Legal**: A dedicated multi-turn **Legal QA** (He et al., 2023) dataset, representing scenarios where users seek professional legal advice.

- **Sports & Entertainment**: Sourced from the **NaturalConv** (Wang et al., 2021) dataset, this domain captures interest-driven discussions regarding sports events, celebrities, and entertainment news.

- **Technology & Education**: Also sourced from **NaturalConv**, this subset focuses on knowledge-sharing interactions related to tech trends, educational topics, and scientific discussions.