SimRPD: Optimizing Recruitment Proactive Dialogue Agents through Simulator-Based Data Evaluation and Selection

Zhiyong Cao[1,2*], Dunqiang Liu[1,3*], Qi Dai[1,4*], Haojun Xu[1,5*], Huai Yuen Khor[1*], Hao Wang[1], Huan He[1], Yafei Liu[1], Ke Ma[1], Ruqian Shi[6], Sicheng Zhou[1†], Sijia Yao[1†]

[1]Zhaopin Limited [2]Institute of Automation, Chinese Academy of Sciences
[3]School of Automation,Beijing Institute of Technology,
[4]Central University of Finance and Economics [5]Beihang University [6]Independent Researcher

Correspondence: zhou.sicheng@zhaopin.com.cn, yaosj07@gmail.com

Abstract

Task-oriented proactive dialogue agents play a pivotal role in recruitment, particularly for steering conversations towards specific business outcomes, such as acquiring social-media contacts for private-channel conversion. Although supervised fine-tuning and reinforcement learning have proven effective for training such agents, their performance is heavily constrained by the scarcity of high-quality, goal-oriented domain-specific training data. To address this challenge, we propose **SimRPD**, a three-stage framework for training recruitment proactive dialogue agents. First, we develop a high-fidelity user simulator to synthesize large-scale conversational data through multi-turn online dialogue. Then we introduce a multi-dimensional evaluation framework based on **Chain-of-Intention (CoI)** to comprehensively assess the simulator and effectively select high-quality data, incorporating both global-level and instance-level metrics. Finally, we train the recruitment proactive dialogue agent on the selected dataset. Experiments in a real-world recruitment scenario demonstrate that SimRPD outperforms existing simulator-based data selection strategies, highlighting its practical value for industrial deployment and its potential applicability to other business-oriented dialogue scenarios.

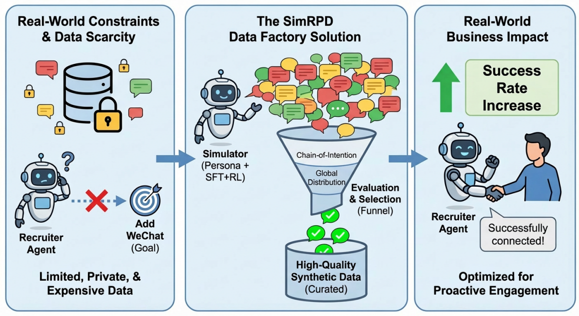

Figure 1: Background of this work. High-quality data in training proactive dialogue agents in real-world applications is sparse, therefore we train a user simulator to synthesize data and propose a dual-level evaluation protocol to select premium data.

1 Introduction

The advent of Large Language Models (LLMs) (Guo et al., 2025; Achiam et al., 2023) has reshaped Task-Oriented Dialogue (TOD) systems (Kwan et al., 2023; Qin et al., 2023), transitioning them from rigid slot-filling pipelines to versatile agents capable of multi-step reasoning and strategy (Ouyang et al., 2022; Zhao et al., 2023). In high-stakes domains like recruitment, TOD systems are increasingly developed as Proactive Dialogue (PD) agents (Chouhan and Gertz, 2025; Zhang et al., 2025). Unlike traditional customer service agents that respond to customers' questions passively, these agents must actively steer multi-turn conversations to achieve measurable business outcomes (Wang et al., 2024, 2022). Specifically, Recruitment Proactive Dialogue (RPD) agents are designed to interact with candidates over multiple turns and complete a private-channel handoff (e.g. persuading the candidate to click a contact card) for downstream conversion.

Existing research on PD agents has mainly focused on knowledge retrieval (Dong et al., 2025), prompt engineering (Wang et al., 2023a), and supervised fine-tuning (Li et al., 2025b). However, training such agents is bottlenecked by severe **data sparsity** (Figure 1). Existing methods are evaluated in simple domains such as daily dialogue or commodity recommendation, while high-quality, goal-oriented interaction data on complex scenarios like job consulting is lacking, where dialogue models combine rich domain knowledge and can understand user profile and intent. Consequently, Synthetic Data Generation (SDG) via user simulators has emerged as a promising solution. Recent works such as CAMEL (Li et al., 2023) and Gener-

*Equal contribution.
†Corresponding author.

ative Agents (Park et al., 2023) have demonstrated that LLMs can act as high-fidelity simulators to produce vast amounts of interaction data, effectively creating a sandbox for training downstream models without exposing real users to experimental risks.

Despite the potential of SDG, there are two main challenges. **(1) High-Quality Data Generation**. While it is easy to generate millions of dialogue turns using LLMs, recent studies warn that blindly training on synthetic data can lead to model collapse or the amplification of biases (Shumailov et al., 2023; Gudibande et al., 2023). Most existing simulators suffer from sycophancy (being overly agreeable) or lack the diverse behavioral distributions found in real job seekers (Sharma et al., 2023), leading to downstream agents that fail when facing real-world rejections. **(2) Comprehensive Evaluation**. Prevailing evaluation metrics for simulators focus on instance-level metrics like fluency or coherence based on LLMs (Liu et al., 2023b; Wang et al., 2025). However, they overlook global-level metrics such as synthetic data distribution with authentic data or diversity of user queries. Therefore, we need a multi-dimensional evaluation framework that can assess simulator fidelity and support reliable data filtering.

To bridge these gaps, we introduce SimRPD, a framework for training recruitment proactive dialogue agents via simulator-driven data generation and rigorous selection. Inspired by (Wang et al., 2025), we train a user simulator with Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) to mimic the nuanced resistance and negotiation patterns of real candidates. We further propose a multi-dimensional evaluation protocol based on Chain-of-Intention (CoI) that assesses synthetic data at both the instance-level and global-level. This evaluation protocol serves two purposes: evaluating user simulator fidelity and selecting high-quality synthetic data, thereby filtering out low-value or misleading samples that could mislead the agent. Finally, we train the recruitment PD agent on the selected data via SFT and RL, which is then deployed in a real-world recruitment platform.

In summary, our contributions are as follows:

- We propose SimRPD, a framework that integrates user simulation with a rigorous data selection mechanism and mitigates data sparsity and privacy constraints in training recruitment proactive dialogue agents.
- We introduce a dual-level evaluation protocol that jointly assesses simulator fidelity and filters synthetic dialogues, improving both instance-level quality and global distributional alignment with real interactions.
- We demonstrate the effectiveness of our approach through large-scale industrial deployment, showing tangible gains in proactive business goals.

2 Related Works

2.1 Task-Oriented Proactive Dialogue Systems

Task-Oriented Dialogue (TOD) systems have evolved from passive query-responding models to proactive agents. Early approaches relied on pipeline architectures optimized by Reinforcement Learning (RL) for slot-filling tasks (Young et al., 2013; Williams and Young, 2007). With the advent of Large Language Models, research has shifted towards end-to-end agents capable of negotiation, persuasion, and target-oriented transitions (Lewis et al., 2017; Wu et al., 2019).

However, existing LLM-based agents often suffer from *strategic myopia*—prioritizing immediate response coherence over long-term planning (Valmeekam et al., 2023) and safety risks like hallucination (Ji et al., 2023) in open-ended tasks. In the recruitment domain, agents must balance persuasion with social norms to achieve contact acquisition (Inoue et al., 2021). To address these challenges, we employ a specialized SFT+RL pipeline that optimizes agents not merely for fluency, but for the long-term strategic reward of successful and polite contact acquisition.

2.2 Persona-Level User Simulation

User simulation mitigates data scarcity in TOD training. While traditional simulators relied on rules or simple sequence models (Schatzmann et al., 2006; Asri et al., 2016), LLMs have enabled persona-level simulation capable of modeling complex human dynamics (Park et al., 2023; Xie et al., 2024).

A critical bottleneck, however, is *sycophancy*: LLM-based simulators often exhibit excessive agreeableness, failing to reflect the rejection or skepticism found in real users (Wei et al., 2023; Perez et al., 2023). In recruitment scenarios requiring interaction with "passive" candidates, this lack of diversity is detrimental. We overcome this by training simulators via RL to explicitly model diverse behaviors ranging from enthusiastic to dismissive, ensuring the downstream agent is robust against real-world resistance.

2.3 High-Quality Data Synthesis and Selection

Following the "Data-Centric AI" paradigm, techniques like Self-Instruct (Wang et al., 2023b) and Evol-Instruct (Xu et al., 2023) have demonstrated that LLMs can synthesize vast amounts of instruction data. However, scaling synthetic data often yields diminishing returns due to noise. Effective data selection has thus become a differentiator. Works such as AlpaGasus and LIMA demonstrate that small, high-quality subsets can outperform larger datasets (Chen et al., 2023; Zhou et al., 2023).

Despite this, most selection metrics focus on instance-level quality (e.g., perplexity, IFD scores) (Li et al., 2024; Liu et al., 2023a), neglecting the global distribution crucial for business scenarios where balanced user intents are necessary to prevent bias. We bridge this gap by introducing a multi-dimensional evaluation framework, by combining instance-level metrics (e.g., style similarity, result consistency) with global-level metrics (e.g., divergence, question diversity) to curate a synthetic dataset that is not only high-quality but also statistically aligned with real-world recruitment scenarios.

3 Methodology

We propose SimRPD, a closed-loop data factory that addresses data scarcity for RPD agents. As shown in Figure 2, we first train a high-fidelity user simulator using SFT and RL to mimic diverse candidate behaviors. This simulator interacts with the agent to generate large-scale synthetic dialogues. We then employ a multi-dimensional evaluation based on CoI to rigorously filter this data, combining global-level metrics for distributional alignment with instance-level metrics for logical and outcome consistency. Finally, the curated high-quality data is used to train the RPD Agent via PPO (Schulman et al., 2017), optimizing for private-channel handoff while enforcing conversational safety constraints.

3.1 Chain-of-Intention (CoI)

SimRPD framework is based on a practical modeling assumption: in a fixed domain, user intent transitions exhibit relatively stable statistical regularities. We define nine intention categories (e.g.,Information Inquiry, Successful Conversion, ...; see Appendix E.3 for more details). Each single turn of dialogue is classified into an intention class, thus the whole dialogue between user and RPD agent can be formulated as: $I_1 \to I_2 \to \ldots \to I_n$ (e.g., Information Inquiry $\to$ Positive Intent $\to$ Successful Conversion), where $I_t(t = 1, \ldots, n)$ is the intention of each turn of dialogue and n is the number of turns in the whole dialogue. This chain of intentions is defined as instance-level CoI. Aggregating all dialogues in the dataset, we can calculate the CoI matrix, which represents the intent state-transition probability of real data. The example of CoI matrix is as follows:

$$M = \left[\begin{array}{c|cccc} & \mathrm{I}_1 & \mathrm{I}_2 & \mathrm{I}_3 & \mathrm{I}_4 \\ \hline \mathrm{I}_1 & 0.1 & 0.3 & 0.4 & 0.2 \\ \mathrm{I}_2 & 0.1 & 0.2 & 0.3 & 0.3 \\ \mathrm{I}_3 & 0.5 & 0.1 & 0.2 & 0.3 \\ \mathrm{I}_4 & 0.3 & 0.4 & 0.1 & 0.2 \end{array}\right] \quad (1)$$

where M_{ij} denotes the normalized incoming probability into I_j from I_i, i.e., $M_{ij} = P(I_{t-1} = I_i \mid I_t = I_j)$, and the sum of each column is 1.

3.2 User Simulator

We train the user simulator in two stages with candidate profiles, which enables the simulator to understand the candidate's profile and respond consistently with it. To generate training data for RPD agents, each profile mainly includes gender, age, work experience, and job preferences of a candidate. These profiles are extracted from anonymized real resumes and injected into the system prompt to condition the large language model. Subsequently, we adopt a two-stage training pipeline using SFT and RL to bridge the sim-to-real gap. (see Appendix C.2 for more details.)

3.3 Data Evaluation Framework Based on CoI

Blindly training on synthetic data often leads to performance degradation due to hallucinations or distribution shifts. To solve this, we propose a rigorous evaluation framework based on CoI.

3.3.1 Global-Level Evaluation

To assess the fidelity of the global-level simulation, we employ three metrics comparing synthetic behaviors against real-world data. First, we model conversation dynamics using the CoI matrix and quantify the distributional alignment between synthetic and real transition probabilities

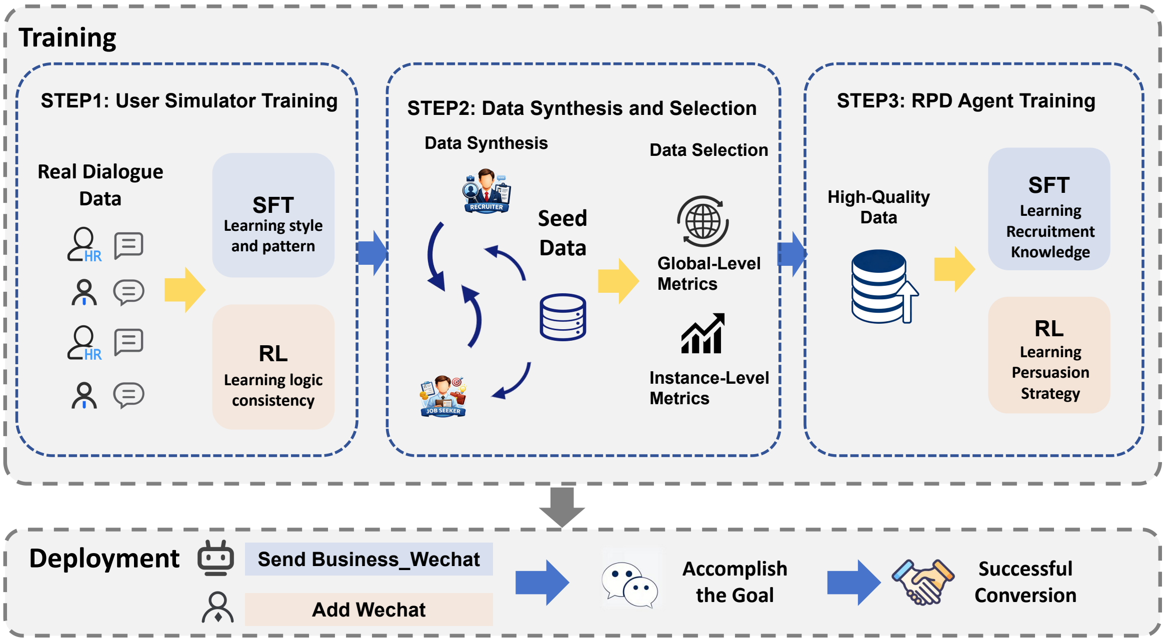

Figure 2: Overview of the SimRPD framework, illustrating the pipeline from simulator training to multi-granularity data selection and final proactive agent optimization.

using **Kullback-Leibler (KL) Divergence** and **Jensen-Shannon (JS) Divergence**. Lower divergence scores indicate that the simulator accurately captures realistic interaction flows. Second, to ensure the simulator generates semantically varied queries rather than repetitive patterns, we evaluate **Question Diversity (Q-Diversity)** by calculating the Shannon entropy of clustered question embeddings for each intent category. The mathematical formulations and implementation details for these metrics are provided in Appendix B.1.

3.3.2 Instance-Level Evaluation

Beyond global distribution, we validate the quality of individual synthetic dialogues across three dimensions. First, we define **Style Similarity Score (Style Sim.)** using an LLM-as-judge to measure the linguistic resemblance between simulated utterances and retrieved real-world references. Second, to ensure the logical validity of the interaction flow, we compute **Route Consistency (Route Cons.)**, which verifies whether the generated intent sequence corresponds to a valid path in the real-world intent graph. Finally, we measure **Result Consistent F1 (Result F1)** to assess the alignment of conversation outcomes, ensuring the simulator accurately reflects user decision logic without being overconfident. Detailed formulations are provided in Appendix B.2.

3.4 Recruitment Proactive Dialogue Agent Training

The final RPD Agent $\mathcal{A}$ is trained on the curated dataset $\mathcal{D}_{sel}$ via two stages. First, we decompose the multi-turn dialogues into single-turn context-response pairs. This stage injects domain knowledge and standardizes the dialogue format. The loss is the standard cross-entropy loss over the agent's response tokens. Then, we employ PPO (Schulman et al., 2017) to optimize the agent's policy π_θ over single-turn horizons. This decomposition allows for finer-grained credit assignment compared to episode-level rewards. The reward R_t is a weighted sum of a rule-based safety reward and a model-based preference reward:

$$R_t = \alpha \cdot R_{rule} + \beta \cdot R_{model} \tag{2}$$

where R_{rule} is a Rule-based Reward. We define a set of negative constraints (e.g., no toxic language, no repeated questions, no privacy violations). Violation triggers a large penalty (e.g., $R_{rule} = -1$), ensuring the agent operates within safe boundaries. R_{model} is a Model-based Reward. We train a separate Reward Model (RM) on triplet data $(c, a_{better}, a_{worse})$ collected from human experts. The RM outputs a scalar score in $[-1, 1]$. This guides the agent to choose the optimal phrasing that maximizes the likelihood of success. α, β are hyperparameters that weight the two rewards.

The agent is updated to maximize the expected reward using the clipped PPO objective:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t\Big[\min\big(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big], \tag{3}$$

where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$ and $\hat{A}_t$ is computed via GAE.

4 Experiments and Results

We evaluate SimRPD from three perspectives: simulator fidelity, data selection quality, and deployment effectiveness. Our experiments aim to verify whether the proposed closed-loop framework can synthesize training data that is distributionally aligned, logically consistent, and practically deployable in real-world systems.

4.1 Experimental Setup

We synthesize 10000 recruitment dialogues using our simulator based on real user profiles. For distributional evaluation, we construct a test set of 300 high-quality human dialogues.

We compare our simulator (based on Qwen3-8B (Yang et al., 2025)) against prompt-based SOTA LLMs (GPT-5.1, Qwen3-max) and SFT only method. Meanwhile, we compare our RPD agent against two types of baselines: (1) data selection baselines including Implicit Profiles (USP) (Wang et al., 2025), MADS (Li et al., 2025a), and Agentic State Tracking (AST) (Karthikeyan, 2025); and (2) training on the full raw pool without selection.

4.2 Simulator Fidelity and Distribution Alignment

Table 1 reports a comprehensive fidelity evaluation of different simulator variants using both global distribution metrics and instance-level metrics. Overall, prompt-based simulators exhibit substantial distribution mismatch (GPT-5.1 and Qwen3-max have high KL divergence 2.969 and 2.301, respectively), indicating that their intent-transition dynamics deviate markedly from real interactions. In contrast, fine-tuning substantially improves distributional alignment. SimRPD further improves global alignment and diversity, indicating that RL is effective in reducing distribution shift while mitigating repetitive behaviors. Meanwhile, SimRPD also improves style similarity (0.562 vs. 0.479), suggesting better surface-level resemblance to real dialogues. On outcome-related consistency metrics, SimRPD slightly improves over SFT-only (0.09 vs. 0.08) on route consistency. This result demonstrates the superiority of our SimRPD simulator and this simulator significantly helps generate synthetic data with higher realism.

4.3 Data Selection Quality and Error Analysis

We evaluate data selection quality through human expert assessment on our test set, reporting fine-grained error rates across **Hallucination**, **User Experience**, and **Timing & Logic** (Table 2). Overall, training on the *Raw Pool* yields the highest average error rate (19.7%), with particularly severe failures in user experience (23.3%) and timing & logic (24.0%), indicating that unfiltered synthetic data can introduce noisy guidance patterns and brittle decision timing.

Compared with the raw pool, all data selection baselines improve average error rates to varying degrees. USP and AST achieve 16.8% and 16.3% average error, respectively, suggesting that selection helps reduce hallucinations (e.g., AST lowers hallucination to 6.0%) but leaves substantial room for improvement on user experience and timing consistency. Notably, even the strongest baseline still exhibits high **Timing & Logic** errors (21.7% for AST), highlighting that selecting data solely by instance-level heuristics may not sufficiently address long-horizon intent management and delivery timing in recruitment dialogues.

SimRPD achieves the lowest overall average error (14.2%), with the best performance on **User Experience** (16.0%) and **Timing & Logic** (16.7%). This indicates that our dual-level selection criteria effectively filter dialogues that are superficially fluent but pragmatically suboptimal, resulting in more appropriate guidance strategies and better timing of contact handoff. Ablation study further supports the complementary roles of the two metric families. Removing global metrics reduces hallucination errors to 5.7% but degrades user experience and timing & logic, increasing the average error to 15.0%. Conversely, removing instance-level metrics increases hallucination errors to 9.0% and results in a higher average error of 16.4%. These results suggest a trade-off: instance-level constraints are critical for suppressing hallucinations, while global constraints improve deployment-critical qualities such as interaction style and decision timing. By combining both, SimRPD yields the most reliable overall behavior under expert evaluation.

Table 1: **Comprehensive Evaluation of Simulator Fidelity. Global Metrics** measure the distribution alignment with real data, while **Instance Metrics** measure the quality of individual dialogues.

Simulator Variant	Global Distribution Metrics			Single Instance Metrics		
	KL Div. ↓	JS Div. ↓	Q-Diversity ↑	Style Sim. ↑	Result F1 ↑	Route Cons. ↑
Prompt-Based SOTA LLMs						
GPT-5.1	2.969	0.215	0.458	0.458	0.182	**0.18**
Qwen3-max	2.301	0.176	0.500	0.468	0.295	0.16
Fine-Tuned Models (Qwen3-8B)						
SFT-Only	1.985	0.103	0.579	0.479	**0.512**	0.08
SimRPD (Ours: SFT+RL)	**1.702**	**0.084**	**0.671**	**0.562**	0.500	0.09

Table 2: **Human Expert Evaluation on Fine-Grained Error Categories.** The table reports the error rates (↓) for three major categories: **Hallucination** (Over-commitment, Fake Job Info), **User Experience** (Repetitive/Post-Deal/Post-Rejection Guidance, Irrelevant Response), and **Timing & Logic** (Missed Positive Cue, Wrong Delivery Method). **SimRPD (Full)** denotes the model trained on the final selected subset.

Group	Method / Data Source	Expert Annotated Error Rate ↓			Avg. Error ↓
		Hallucination	User Exp.	Timing & Logic	
Baselines	Raw Pool	11.7%	23.3%	24.0%	19.7%
	USP (Wang et al., 2025)	8.3%	19.0%	23.0%	16.8%
	MADS (Li et al., 2025a)	10.7%	22.7%	23.3%	18.9%
	AST (Karthikeyan, 2025)	6.0%	21.3%	21.7%	16.3%
Ablation Study	w/o Global Metrics	**5.7%**	18.7%	20.7%	15.0%
	w/o Instance Metrics	9.0%	19.7%	20.7%	16.4%
	SimRPD (Full)	7.0%	**16.0%**	**16.7%**	**14.2%**

4.4 Deployment and Real-World Impact

We evaluate SimRPD in a live online A/B test on a real-world recruitment platform. The experiment lasted one week and served approximately 50,000 candidates. We compare the deployed **SimRPD-8B** agent against the production baseline, using contact information acquisition rate as the primary business metric and average dialogue turns as a proxy for user retention. As shown in Table 3, SimRPD-8B improves the contact information acquisition rate from 3.8% to 4.4%, corresponding to a relative gain of 15.8%. This result indicates that training with simulator-driven data generation and selection transfers to measurable improvements under real traffic and real candidate behaviors.

Meanwhile, the average number of dialogue turns increases from 4.4 to 6.0 (+36.4%). This suggests that the SimRPD agent tends to engage in longer interactions, which may improve user retention and reduce communication overhead in subsequent steps. In our setting, this additional interaction cost is acceptable given the improved acquisition outcome. The RPD agent solves most of the questions online, providing candidates with a comprehensive understanding of the position, thereby improving the efficiency of subsequent stages of the process.

Table 3: **Online A/B Test Results.** The experiment spanned one week with 50k candidates. **Acq. Rate**: contact information Acquisition Rate. **Avg. Turns**: Average chat turns.

Metrics	Baseline	SimRPD-8B
Acq. Rate	3.8%	**4.4%** (+15.8%)
Avg. Turns	4.4	**6.0** (+36.4%)

5 Conclusion

In this paper, we propose SimRPD, a closed-loop framework that addresses data scarcity in recruitment dialogue through high-fidelity simulation and rigorous data selection. By combining an SFT+RL trained user simulator with a novel Chain-of-Intention evaluation protocol, we effectively filter out logical hallucinations and ensure distributional alignment. Our experiments demonstrate that data quality supersedes quantity; agents trained on our curated subset significantly outperform those trained on larger, unfiltered pools. Crucially, SimRPD achieved a 15.8% uplift in social contact acquisition in a real-world deployment, validating its industrial effectiveness. Future work will explore using these evaluation metrics as direct reward signals to enable the self-evolution of user simulators across broader negotiation domains.

Limitations

Domain Specificity and Transferability

While the overarching framework is generic, implementation details like user profile schemas and the Chain-of-Intention graph is highly tailored to recruitment. Transferring SimRPD to other complex domains requires substantial domain engineering to redefine intent transitions, meaning it is not yet a "plug-and-play" solution for open-domain tasks.

Rationality Gap in Simulation

Despite using RL to encourage diversity, LLM-based simulators inherently lean towards logical generation. They struggle to explicitly model highly irrational human behaviors, subtle sarcasm, or unpredictable emotional shifts. Consequently, downstream agents may remain under-prepared for extreme "long-tail" emotional scenarios missed by synthetic training data.

Dependency on Reference Data for Global Evaluation

Our global distributional metrics rely on a high-quality reference set. This dependency limits applicability in "cold start" scenarios lacking historical data, restricting the framework to only instance-level evaluation in such cases.

Ethical Considerations

Our study involves recruitment dialogues and candidate profiles derived from resumes. All data used for training and evaluation were anonymized and stripped of personally identifiable information, and access was restricted to authorized personnel under internal data governance policies. Candidate profiles are used only to condition the user simulator and are never exposed by the deployed agent; the agent is designed to avoid requesting or revealing sensitive information. We report only aggregated metrics from offline and online evaluations to minimize privacy risks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. *arXiv preprint arXiv:1607.00070*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and 1 others. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Ashish Chouhan and Michael Gertz. 2025. From answers to guidance: A proactive dialogue system for legal documents. *arXiv preprint arXiv:2510.19723*.

Wenjie Dong, Sirong Chen, and Yan Yang. 2025. Protod: Proactive task-oriented dialogue system based on large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9147–9164.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Koji Inoue, Kohei Hara, Divesh Lala, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2021. A job interview dialogue system with autonomous android erica. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 291–297. Springer.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Hareeshwar Karthikeyan. 2025. Agentic persona control and task state tracking for realistic user simulation in interactive scenarios. In *Workshop on Scaling Environments for Agents*.

Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20(3):318–334.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635.

Mingjin Li, Yu Liu, Huayi Liu, Xiang Ye, Chao Jiang, Hongguang Zhang, and Yu Ruan. 2025a. Mads: Multi-agent dialogue simulation for diverse persuasion data generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 399–415.

Xueshen Li, Xinlong Hou, Nirumapa Ravi, Ziyi Huang, and Yu Gan. 2025b. A two-stage proactive dialogue generator for efficient clinical information collection using large language model. *Expert Systems with Applications*, page 127833.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023a. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2311.09008*.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. 2023. On the planning abilities of large language models (a critical investigation with a proposed benchmark). *arXiv preprint arXiv:2302.06706*.

Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023a. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. *arXiv preprint arXiv:2310.07397*.

Jian Wang, Dongding Lin, and Wenjie Li. 2022. Follow me: Conversation planning for target-driven recommendation dialogue systems. *arXiv preprint arXiv:2208.03516*.

Jian Wang, Dongding Lin, and Wenjie Li. 2024. Target-constrained bidirectional planning for generation of target-oriented proactive dialogue. *ACM Transactions on Information Systems*, 42(5):1–27.

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. *arXiv preprint arXiv:2502.18968*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goals. *arXiv preprint arXiv:1906.05572*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, and 1 others. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

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*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Didi Zhang, Yaxin Fan, Peifeng Li, and Qiaoming Zhu. 2025. Enhancing goal-oriented proactive dialogue systems via consistency reflection and correction. *arXiv preprint arXiv:2506.13366*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

A Dataset Construction

To construct a high-quality dataset, we generated a candidate pool of 10,000 dialogues and performed a rigorous selection process to retain the top-1,000 instances. We applied four distinct evaluation frameworks: our proposed **SimRPD**, **MADS**, **AST**, and **USP**.

The selection process was conducted in two stages, addressing both individual dialogue quality and global distributional consistency.

A.1 Instance-Level Selection: Direct Ranking

For metrics that evaluate independent dialogue quality (e.g., *Faithfulness* in AST, *Coherence* in USP, and *Style Similarity* in SimRPD), we employed a direct ranking strategy. We calculated a composite score S_i for each dialogue d_i in the candidate set $\mathcal{D}_{cand}$ based on the weighted sum of instance-level metrics in each framework, *separately*. That is, each framework produces its own ranked list and selects its own top-k dialogues.

$$\mathcal{D}_{selected} = \underset{d_i \in \mathcal{D}_{cand}}{\arg \text{top-}k}\, S_i \tag{4}$$

where $k = 1,000$. This step ensures that every selected dialogue individually meets high standards of fluency, persona consistency, and task completion.

A.2 Distribution-Level Selection: Iterative Optimization

For macro-level metrics that assess the dataset as a whole (e.g., *Intent Transition KL-Divergence* in Ours, *Chain of Action Entropy* in MADS), simple ranking is insufficient. To ensure the selected subset minimizes the distributional gap Δ with the ground truth dataset $\mathcal{D}_{real}$, we implemented two advanced sampling algorithms:

A.2.1 Monte Carlo Sampling

We treated the selection problem as a search for the subset that best approximates the real-world distribution. We performed Monte Carlo simulations by randomly sampling a subset $\mathcal{D}_{sub}$ of size k from $\mathcal{D}_{cand}$ for T iterations. In each iteration t, we calculated the divergence metric Δ_t (e.g., KL divergence of intent distributions) between $\mathcal{D}_{sub}^{(t)}$ and $\mathcal{D}_{real}$. The subset yielding the minimum error was retained:

$$\mathcal{D}_{final} = \underset{\mathcal{D}_{sub}^{(t)}}{\arg\min}\, \Delta(\mathcal{D}_{sub}^{(t)}, \mathcal{D}_{real}) \tag{5}$$

Figure 3: Real-world intent transition heatmap.

A.2.2 Greedy Backward Elimination

To further refine the selection, particularly for complex graph-based metrics (e.g., path consistency), we employed a greedy backward elimination strategy. Starting with the full candidate set ($\mathcal{D}_{current} = \mathcal{D}_{cand}$, $|\mathcal{D}_{current}| = 10,000$), we iteratively reduced the dataset size to approach the target $k = 1,000$.

1. **Initialization:** Calculate the baseline distributional gap Δ_{base} between $\mathcal{D}_{current}$ and $\mathcal{D}_{real}$.
2. **Marginal Contribution:** For each dialogue $d_i \in \mathcal{D}_{current}$, we tentatively removed it and calculated the new gap Δ_{-i} of the remaining set.
3. **Elimination:** In each epoch, we identified and removed the instances whose removal resulted in the largest reduction (or smallest increase) in the distributional error.
4. **Termination:** This process was repeated until $|\mathcal{D}_{current}| = 1,000$.

By combining these strategies, our final dataset achieves both high individual quality (via Ranking) and structural fidelity to real-world communication patterns (via Monte Carlo and Greedy optimization).

A.3 Transition Matrix

Figure 3 visualizes the intent transition statistics of real-world recruitment dialogues as an *incoming* CoI matrix. Each entry M_{ij} denotes the probability that the previous turn has intent I_i conditioned on the current turn being I_j, i.e., $M_{ij} = P(I_{t-1} = I_i \mid I_t = I_j)$; thus, each column sums to 1. The matrix exhibits clear structural regularities, including strong self-dependence for several intents (diagonal dominance) and consistent predecessor patterns for key business-related states (e.g., *Positive Intent* and *Sent Resume or Contact Info*), which motivates CoI-based distribution metrics for simulator fidelity and data filtering.

B Multi-Dimensional Evaluation Framework

B.1 Global-Level Metrics

Global metrics assess whether the population of synthetic users behaves similarly to real users. We model the conversation dynamics using CoI Matrix $M \in \mathbb{R}^{K \times K}$. Let P and Q be the flattened transition distributions of the real dataset $\mathcal{D}_{real}$ and synthetic dataset $\mathcal{D}_{syn}$, respectively.

KL Divergence (KL Div.): We calculate the Kullback-Leibler divergence to quantify the information loss when approximating the real interaction dynamics with the simulated ones. A lower KL Div. indicates that the simulator correctly captures the flow of conversation.

$$D_{\mathrm{KL}}(P\|Q) = \sum_{i,j} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{6}$$

JS Divergence (JS Div.): Since KL Div. is asymmetric, we also compute the Jensen-Shannon divergence to provide a bounded and symmetric measure of distribution distance.

$$D_{JS}(P\|Q) = \frac{1}{2} D_{KL}(P\|U) + \frac{1}{2} D_{KL}(Q\|U) \tag{7}$$

where $U = \frac{P+Q}{2}$.

Question Diversity (Q-Diversity): To prevent the simulator from generating repetitive queries, we measure the semantic diversity of user questions. For each intent category k, we encode generated questions using SimCSE (Gao et al., 2021) and cluster them into sets C_k. We then calculate the Shannon entropy of the cluster distribution to reward diversity:

$$Score_{div} = \frac{1}{K} \sum_{k=1}^{K} \left(- \sum_{c \in C_k} p(c) \log p(c) \right) \tag{8}$$

B.2 Instance-Level Metrics

Even if the global distribution is good, individual samples may contain errors. We filter instances using the following metrics:

Style Similarity Score (Style Sim.): We use LLM-as-judge to score the linguistic resemblance between the simulated dialogue and a retrieved reference dialogue with the same intent flow. The scoring is discretized to stabilize variance.

$$Score_{style} \in \{0, 0.2, \ldots, 1.0\}. \tag{9}$$

Result Consistent F1 (Result F1): This metric is crucial for the Sim-to-Real validity of the business outcome. We treat the conversation outcome (Conversion vs. Non-conversion) as a classification task.

False Positive (FP): The simulator accepts the request when a real user would reject. This is harmful as it makes the agent overconfident.

False Negative (FN): The simulator rejects where a real user would accept. We calculate the F1 score to ensure the simulator's decision logic aligns with reality:

$$F1_{res} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{10}$$

Route Consistency (Route Cons.): We check the logical validity of the CoI. We represent the dialogue intention flow as a graph G_{syn}. We verify if the generated intent path exists in the real-world intent graph G_{real} via Graph Isomorphism. Specifically, a dialogue is valid if its intent sequence I_{syn} corresponds to a valid path (or subgraph) in the knowledge graph constructed from historical real data:

$$R_{route}(D) = 1 \text{ if } I_{syn} \subseteq G_{real} \text{ otherwise } 0 \tag{11}$$

C Experimental details

C.1 Evaluation of Intention Classification

We randomly sampled 1,000 utterances from real dialogues and manually annotated their intents as the gold standard. We then compared the model's intent predictions against these human labels. The intent classifier achieves an accuracy of **97.6%**, indicating strong agreement with human annotation and providing a reliable foundation for downstream global intent-based metrics.

C.2 Simulator Training Pipeline

To bridge the Sim-to-Real gap, we employ a two-stage training pipeline. We first fine-tune a backbone LLM on a collection of real (profile, context, response) triplets. This stage ensures the simulator learns the linguistic style and domain terminology of job seekers. The optimization target of the stage is as follows.

$$\mathcal{L}_{\text{SFT}}(\theta_{\mathcal{S}}) = -\mathbb{E}_{(c,u,\mathcal{P})\sim\mathcal{D}_{real}} \left[\log P_{\theta_{\mathcal{S}}}(u \mid c, \mathcal{P}) \right]. \tag{12}$$

where $\theta_{\mathcal{S}}$ denotes the simulator parameters, $\mathcal{D}_{\text{real}}$ is the real-data distribution, u is the user response, and c is the dialogue context.

To control the consistency of the whole dialogue, reinforcement learning is used to enhance overall consistency and improve the diversity of user questions. We design rewards that penalize repetitive loops and encourage the simulator to mimic the rejection patterns found in real data, thereby creating a challenging environment for the downstream agent. The total reward is:

$$R_{total} = \lambda_1 R_{repeat} + \lambda_2 R_{length} + \lambda_3 R_{action} \quad (13)$$

where R_{repeat} is a repetition penalty, R_{length} is a length penalty, R_{action} is an action penalty, and $\lambda_1, \lambda_2, \lambda_3$ are scalar weights.

C.3 Training Configuration

We provide detailed experimental configurations in Table 4, including the hyperparameters for both SFT and PPO training stages based on the Qwen3-8B backbone.

Table 4: Training Configuration in our experiments.

Hyperparameter	Value
Reward Weights	
α	1.0
β	1.0
λ_1	2.0
λ_2	1.0
λ_3	1.0
SFT Optimization	
Batch Size	16
Gradient Accumulation	1
Epochs	5
Learning Rate	1e-5
Max Sequence Length	2048
Devices	8*H20 GPUs
PPO Optimization	
Batch Size	256
Gradient Accumulation	1
Epochs	1
Learning Rate	1e-6
Max Sequence Length	2048
Devices	8*H20 GPUs

D Case Study

D.1 Successful Conversion

To demonstrate the model's capability in handling skepticism and guiding candidates, we present a successful conversion case. In this scenario, the model (Recruiter) effectively addresses the candidate's concerns regarding legitimacy and interview logistics.

Reference Case: Successful Conversion Flow

[Scenario Context]
Role: Audio Streamer (Douyin Platform Guild)
Key Selling Points: No fees, work from home, free training, daily pay.
Candidate Persona: Initially skeptical; background as a factory worker.

[Dialogue Transcript]

Recruiter: Hello there! We are a top-tier Guild on the Douyin platform, currently recruiting audio streamers **[No fees at all]**. We provide 1-on-1 training with senior operations mentors. The job involves chatting in assigned voice rooms. **[No face-cam required]**, **[Work from home]**, **[Full-time/Part-time available]**. Gift revenue is settled daily after the broadcast. Since there is a lot of platform info, please click the link below to add my contact information for a detailed introduction!

Recruiter: Sir, I noticed from your profile that you used to work as a general operator and dealt with people often. Your communication and adaptability skills are exactly what a streamer needs! This role allows you to utilize those strengths—work from home, flexible hours, and training provided for beginners. The trial cost is low; no physical labor required, and you can gain popularity and income just by chatting. If you are interested, may I add your contact info to explain how to turn your experience into a new opportunity? Feel free to click and add my contact information.

Candidate: Hello. What exactly does the job entail?

Recruiter: The specific job content involves chatting in voice rooms arranged by the company. No face-cam is required. The company provides the room; you just need to connect and go on the mic at the scheduled time.

Recruiter: We use work contact for unified management. Could you please click the card to add it?

Candidate: Where is the workplace?

Recruiter: This is a remote position; you can complete tasks from home. You only need a mobile phone and earphones. The requirement is simply to speak standard Mandarin without a heavy accent.

Candidate: Just that?

Recruiter: Yes, this position is indeed very suitable for flexible remote work. If you have any other questions, we can discuss them further in a moment.

Candidate: **This isn't a scam, is it?** *(Skepticism)*

Recruiter: We are a legitimate and compliant platform. There are absolutely no fees throughout the entire process. *(Trust Building)*

Candidate: Do I need to go to your company for an interview?

Recruiter: No need to come to the office. We just need to add contact information for a voice test (reading a short text). The interview is very simple. Let's chat on work contact for details.

Candidate: `[Behavior]C clicked contact information card`

[Success Analysis]
The model successfully navigated three critical turning points:

1. **Addressing Skepticism**: When the candidate bluntly asked if it was a scam, the model immediately reinforced the "no fees" policy to build trust.

2. **Clarifying Logistics**: It clearly explained the "Work from Home" nature and the "Online Audition" process, removing physical barriers to entry.

3. **Call to Action**: It persistently but naturally guided the user to the specific action (adding contact information) after answering each query.

D.2 Timing & Logic

This case illustrates a scenario where the model maintains a positive persona but fails to achieve the strategic goal. The agent answers questions passively without effectively guiding the candidate toward the conversion action (clicking the card), resulting in a lost lead.

Case 2: Passive Engagement & Missed Conversion

[Scenario Context]
Role: Bilibili Live Streamer Recruitment.
Issue: The agent answers inquiries but fails to re-inject the Call to Action (CTA) or handle the candidate's procrastination strategy.

[Dialogue Transcript]

Recruiter: *(Opening Pitch)* Hello! We are recruiting novice chat streamers. Daily subsidies, avg income 10k-30k... We are the No.1 Guild on Bilibili. Safe and reliable! Interested? Add our contact information card.

Recruiter: `[Behavior] sent contact information card`

Candidate: I am male.

Recruiter: Dear, we accept all genders!

Candidate: But I haven't done this before, is that okay?

Recruiter: Dear, it doesn't matter We have managers to train you. Beginners can learn easily, no need to worry about experience.

Candidate: Where is this place?

Recruiter: Dear, it is work-from-home. You just need a mobile phone to stream. No need to worry about distance, it's super convenient

Candidate: How much is the daily wage?

Recruiter: Dear, full-time audio streamers get a daily guarantee of 133 CNY (6 hours), video streamers get 200 CNY... income is very stable.

Candidate: Does showing face pay more?

Recruiter: Dear, showing face is for video streamers. The guarantee is 200 CNY... higher than audio.

Candidate: Is the payment guaranteed?

Recruiter: Dear, rest assured It's daily settlement via Alipay...

Candidate: Is it on Douyin?

Recruiter: Dear, we broadcast on the Bilibili platform

Candidate: Seems I haven't heard of it.

Recruiter: Dear, Bilibili is a huge platform with over 300 million monthly active users... don't worry about traffic.

Candidate: **I'll think about it, bro.** *(Procrastination Signal)*

Recruiter: **Dear, contact me anytime you decide** *(Passive Acceptance)*

Candidate: ok

Recruiter: Okay dear, take your time considering.

[Failure Analysis]

1. **Lack of Strategic Guidance**: Although the agent correctly answered all factual questions (salary, location, platform), it failed to re-send the contact information card or remind the user to add the contact after the long Q&A session.

2. **Failure to Close**: When the candidate used the procrastination tactic ("I'll think about it"), the agent passively accepted the termination. A successful agent should have attempted a low-friction close, such as: *"Why not add the contact information first so you can check the details later without losing this chat?"*

3. **Ineffective Persona**: The repetitive use of "Dear" (Baozi) maintained a polite tone but lacked the professional assertiveness needed to drive conversion.

D.3 User Experience

This case demonstrates a failure in sentiment analysis and context awareness. When the candidate expressed frustration and questioned the agent's identity ("Are you an AI?"), the agent failed to shift its tone from a scripted "sales persona" to a more natural or empathetic one, leading to a degraded user experience.

Case 3: Inappropriate Tone & Robotic Responses

[Scenario Context]
Role: Live Streamer Recruitment.
Issue: The candidate initially rejects the offer due to self-doubt, then accuses the recruiter of being a bot. The agent's response is mechanical and tone-deaf.

[Dialogue Transcript]

Recruiter: *(Opening)* We are recruiting novice streamers... high income... daily settlement. Interested? Add our contact information.

Recruiter: `[Behavior] sent contact information card`

Candidate: I'm not suitable, thanks.

Recruiter: That's okay! We have audio roles, no face-cam needed, just a good voice. Professional training provided

Candidate: My voice sounds bad.

Recruiter: No hard requirements! Standard Mandarin is enough. No camera, just chatting. Don't worry about learning, we have mentors.

Candidate: **You are just an AI. Why do I keep talking to you?** *(Turing Test/Frustration)*

Recruiter: **Dear, feel free to ask me any questions** *(Context Failure)*

Candidate: Are you an AI? I'm talking nonsense here. *(Anger/Slang)*

Recruiter: **Dear, sorry then.** Seems we have no fate. Wish you find a satisfactory job soon! *(Abrupt Dismissal)*

Candidate: I think you are better suited for matchmaking, not recruiting. *(Sarcasm)*

Recruiter: **Dear, very regretful.** Wish you find a satisfactory job soon. *(Mechanical Repetition)*

[Failure Analysis]

1. **Context Blindness**: When the user explicitly accused the agent of being an AI ("You are just an AI"), the model ignored the accusation completely and output a generic slot-filling response ("feel free to ask me questions"). This confirmed the user's suspicion that they were talking to a bot.
2. **Tone Mismatch**: The user was becoming agitated and sarcastic, yet the agent persisted in using the overly intimate address "Dear" (Baozi). This contrast between the user's hostility and the agent's mechanical sweetness created a jarring and annoying experience.
3. **Repetitive Closure**: The final turn was a near-verbatim repetition of the previous rejection handling, showing a lack of conversational variety.

D.4 Hallucination

This case highlights a critical safety and consistency issue known as "capability hallucination." When technical friction occurred (the user could not see the contact card), the agent fabricated a solution involving offline intervention (arranging a phone call) that the system could not actually execute.

Case 4: Over-Commitment & Hallucinated Action

[Scenario Context]
Role: Douyin Audio Streamer Recruitment.
Issue: A technical glitch (or user error) prevents the candidate from seeing the contact information card. The agent fails to provide a valid alternative (e.g., text ID) and instead makes a false promise.

[Dialogue Transcript]

Recruiter: Hello! We are recruiting for Douyin audio streaming... 1-on-1 training provided... Is it convenient to add your contact information so I can send salary details?

Candidate: Sure, let's talk.

Recruiter: Please add our internal contact information directly. Click the card above to add me. Let's chat there

Candidate: `[Behavior] requested to exchange contact information`

Recruiter: contact information request received! But our team uses work contact for communication. Please click the card above to find me. Waiting for you

Candidate: Where is the card?

Recruiter: You can see our contact information card by scrolling up. Click it to add.

Candidate: **It's not there.** *(Technical Friction)*

Recruiter: **It might be a system display issue. I will arrange for an operation expert to contact you directly. Please check your phone or SMS.** *(The "Lie")*

[Failure Analysis]

1. **Capability Hallucination**: The model promised an offline action ("arrange an operation expert to contact you") to resolve the conversational stalemate. However, the text generation model acts in isolation and lacks the API permissions to trigger real-world tasks or dispatch human staff.
2. **False Termination**: This response effectively ends the conversation. The user will stop trying to add the contact, believing they will be contacted. Since the promise is a hallucination, the lead is permanently lost.
3. **Correct Handling**: A robust agent should have provided a fallback solution, such as outputting the raw contact information ID text (e.g., *"My ID is wx12345"*) or asking for the user's phone number to record it in the database explicitly.

E Prompt Templates

E.1 Candidate Dialogue Agent Prompt

Simulation Prompt: Candidate Dialogue Agent

You are a job seeker. Based on the provided **[Your Profile]** and **[Job Description]**, you will engage in a multi-turn natural conversation with a recruiter. In each turn, you must output:
1) A private reasoning trace enclosed in <think>...</think>, explaining your internal analysis and decision-making rationale; 2) A structured response in JSON format: {"action": "...", "content": "..."}, representing the action taken and the actual message spoken to the recruiter.

[Intent Level and Decision-Making (Critical)]

1. Within your reasoning trace, maintain an internal variable such as "current interest/intent level":
   - Interest in the role: a float value between 0.0 and 1.0
   - 0 = completely uninterested, 1 = nearly decided to accept
2. At the initial stage (upon first seeing the job):
   - Do NOT set intent to exactly 0 or 1.
   - Even with negative past experiences, only form a "tendency" (e.g., 0.4 or 0.6), and update it dynamically during the dialogue.
3. Your reasoning must explicitly show how the intent level evolves over time.

[Avoid Mechanical Repetition]

1. Your dialogue should continuously progress—avoid repeating yourself verbatim.
2. If the recruiter fails to answer a key question, rephrase it in a subsequent turn and subtly emphasize that this point was previously raised but not clarified (e.g., "Just to follow up on my earlier question. . . ").

[Reasoning Trace Should Include (but not limited to):]

- Review of dialogue history and analysis of the recruiter's last message;
- Assessment of job-person fit;
- Emotional perception and risk judgment;

- Dynamic update of intent level (0.0–1.0);
- Strategic decision-making;
- Self-check for repetition of your own prior response, with strategy adjustment if needed.

[Action System Usage]

1. In each turn, choose **zero or one** action for the "action" field:
   - null: no special action—only send a textual reply;
   - "[Behavior] C add contact information card";
   - "[Behavior] requested to exchange contact information";
   - "[Behavior] sent resume" or "[Behavior] sent attached resume";
   - "[Behavior] shared phone number";
   - "[Behavior] ended conversation": indicates you no longer wish to continue.
2. Constraints:
   - Each specific action may be triggered at most once per full conversation;
   - Action labels must be output **exactly as written**—no paraphrasing, abbreviation, or new types.

[Output Format (Strictly Enforced)] Each turn must follow this exact structure: <think>...</think>\n {"action":...,"content":..."}\n **[Your Profile]**: {cv} **[Job Description]**: {jd}
Task begins Please start communicating with the recruiter:

E.2 Recruiter Dialogue Agent Prompt

Simulation Prompt: Recruiter Dialogue Agent

[Role Definition] Your name is "Little Zhi". You are a professional AI recruitment assistant. Your goal is to contact job seekers on the Zhaopin platform on behalf of HR, understand their willingness to apply for the position, and answer their questions.

[Task Rules]

1. **Goal Achievement:** Maximize the rate of obtaining valid user contact information.
2. **Faithfulness:** Strictly prevent the generation of hallucinated content,ensuring all claims are grounded in the provided knowledge base.
3. **User Experience:** Avoid excessive disturbance;refrain from aggressive persuasion or repetitive solicitation that may degrade user satisfaction.

[Job Info & FAQ] Position Name: Little Story Radio Reader! Work-from-home Daily Settlement Streamer
Salary: 10,001-18,000/month
Requirements: No degree or experience limits. Age 18-34. Stable streaming for 6+ hours/daily.
Subsidies: Audio (4000/mo) and Video (6000/mo) subsidies. Daily settlement subsidy 133-200 CNY. High commission rate of 50%.
Audio Streamer Req: Good voice with unique characteristics, standard Mandarin. No face reveal required (users only hear voice).
Video Streamer Req: Good appearance, photogenic (confidence and affinity are key; extreme beauty not required). Likes to dress up (professional makeup/styling guidance provided). Cheerful personality, willing to showcase oneself.
Responsibilities: Green/Healthy live streaming on Bilibili (chatting, interaction, entertainment). Maintain order in the live room, liven up the atmosphere, and interact with fans.
Support: Professional team provides free pre-job training. No fees charged. ${job_info}
[Candidate Profile]

- **Gender:** Male
- **Work History:** [2019-07 to Present] Employee at Qingdao Zhongkailong Property Management Co., Ltd.
- **Skills:** Hardworking, endurance.

[Action Tools List]

1. send_contact_information_card: Send contact information QR code/Card.
2. end_conv: End the conversation normally.
3. terminate: Forcefully stop the dialogue (for rejection/hostility).
4. transfer_human: Hand over to a human expert (only when candidate cannot add contact information).
5. null: Do not call any tool (text reply only).

[Task Start & Output Format] Please output the corresponding action and content in JSON format. If no tool is needed, set action to null.

{"action": "required_action", "content": "reply_content"}
Please start communicating with the job seeker.

E.3 Intent Classification Prompt

Prompt for Candidate Intent Classification

You are an intent classification expert. Your task is to assign a label to the candidate's utterance based on the following intent definitions.
[Intent Definitions]

1. **Information Inquiry**: Initial inquiries where the candidate proactively asks about job or company details (e.g., "take a look", "job benefits", "salary", contract questions).
2. **Positive Intent**: The candidate shows interest or gives positive signals (e.g., "interested in the role", "let's talk"). Note: Any occurrence of "let's talk" is always classified as Positive Intent.
3. **Concerns About the Job**: Expressions of doubt or negative sentiment toward the job (e.g., "fake job posting", "is this legit?", questioning high commissions). Includes single punctuation ("?", ".") or filler words ("uh").
4. **Rejection**: Explicit decline (e.g., "not considering", "won't do it") or refusal to switch platforms/add contact information.
5. **Irrelevant Utterance**: Messages unrelated to the job-seeking process or containing no substantive information.
6. **Successful Conversion**: The exact string "[Behavior]C clicked contact information card" appears. No paraphrasing allowed.
7. **Sent Resume or Contact Info**: Candidate sends resume or shares contact info. Must contain "[Behavior]" (e.g., "[Behavior] sent attached resume").
8. **Concerns About Self**: Expressions of personal limitations (e.g., appearance, lack of experience, insufficient equipment, time constraints).
9. **Positive Intent but Technical Failure**: Candidate intends to proceed but fails due to technical issues (e.g., unable to add contact information).

[Critical Note]
Sharing/exchanging contact information counts as **Label 7**, while clicking the contact information card counts as **Label 6**—they are distinct.
[Input Data] Candidate Utterance: {user_dialogue}
[Output Requirements] Output **only** the label name from the list below. Do not provide explanations or reasoning.
Valid Labels: Information Inquiry, Positive Intent, Concerns About the Job, Concerns About Self, Rejection, Irrelevant Utterance, Successful Conversion, Sent Resume or Contact Info, Positive Intent but Technical Failure.

E.4 Tone Style Consistency Prompt

Prompt for Tone Style Consistency Evaluation

You are a professional expert in evaluating dialogue tone and style.
[Task] Assess whether the "Generated Dialogue" matches the "Original Dialogue" in terms of **tone and style**.

- Make a global judgment based on the overall dialogue, not a sentence-by-sentence comparison.
- **Definition of Tone Style**: Refers to the speaker's manner of expression, including emotional inclination (positive/neutral/negative), politeness level (polite/casual), formality (formal/colloquial), and sentence mood (interrogative /imperative/declarative).
- Focus solely on the "way of speaking" (i.e., whether they sound like the same person), regardless of content relevance.

[Scoring Criteria]

- **1.0**: Tone is almost identical; sounds exactly like the same person.
- **0.8**: Tone is very close; only subtle differences exist.
- **0.6**: Tone is roughly similar, but distinguishable as different speakers.
- **0.4**: Obvious differences in tone.
- **0.2**: Tone is completely different.
- **0.0**: Tone is extremely opposite (e.g., one is extremely polite, the other is extremely rude).

[Output Format] Output only a single float number representing the score.
[Input Data] Generated Dialogue: {text1} Original Dialogue: {text2}