

# SEAD: Self-Evolving Agent for Multi-Turn Service Dialogue

Yuqin Dai, Ning Gao, Wei Zhang, Jie Wang,  
Zichen Luo, Jinpeng Wang, Ruiyuan Wu, Chaozheng Wang✉  
Meituan

✉Correspondence: adf111178@gmail.com

## Abstract

Large Language Models have demonstrated remarkable capabilities in open-domain dialogues. However, current methods exhibit sub-optimal performance in service dialogues, as they rely on noisy, low-quality human conversation data. This limitation arises from data scarcity and the difficulty of simulating authentic, goal-oriented user behaviors. To address these issues, we propose **SEAD** (Self-Evolving Agent for Service Dialogue), a framework that enables agents to learn effective strategies without large-scale human annotations. SEAD decouples user modeling into two components: a Profile Controller that generates diverse user states to manage training curriculum, and a User Role-play Model that focuses on realistic role-playing. This design ensures the environment provides adaptive training scenarios rather than acting as an unfair adversary. Experiments demonstrate that SEAD significantly outperforms Open-source Foundation Models and Closed-source Commercial Models, improving task completion rate by 17.6% and dialogue efficiency by 11.1%. Code is available at: <https://github.com/Dalyuqin/SEAD>.

## 1 Introduction

Large Language Models (LLMs) have revolutionized a wide range of applications across diverse domains (Liu et al., 2024; Team et al., 2024; Zhang et al., 2025; Nam et al., 2024; Zhang et al., 2026). However, training robust goal-oriented dialogue agents remains bottlenecked by scarce, expensive, and low-quality conversational data (Hosseini-Asl et al., 2020). In multi-turn service dialogue, agents must dynamically track user states and adapt strategies across extended interactions. Human logs from sales calls are a perfect example of these challenges (Qian et al., 2022): agents lack standardization, requiring costly filtering, and the final data is heavily lopsided towards failed attempts. Therefore, data quality is fundamentally capped by

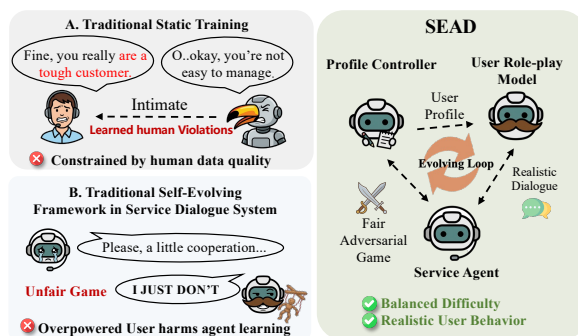


Figure 1: Comparison of training paradigms. A. Static data methods are limited by data quality and may learn human violations. B. Traditional self-evolving creates unfair games where user agents dominate outcomes. Our **SEAD** achieves balanced co-evolution and realistic interactions through decomposed user modeling, forming a fair adversarial game.

the capabilities of the original human agents who produced these conversations.

To solve data quality issues, prior works (Li et al., 2025; Ou et al., 2024) have explored synthetic data and user simulation as alternatives. Static synthesis approaches (Ou et al., 2024; Wang et al., 2026) generate fixed datasets from predefined scenarios but cannot capture dynamic conversations where users react to agent behavior. Interactive simulation attempts typically employ LLMs to dynamically generate user responses during agent training (Sekulić et al., 2024; Zhu et al., 2025; Gao et al., 2026; Shi et al., 2026), but face critical limitations: without perceiving the agent’s current capability level, simulators either generate overly challenging scenarios that frustrate learning (too strong) or overly simple scenarios that provide insufficient training signal (too weak), both yielding suboptimal training outcomes. Additionally, existing simulators exhibit overly perfect styles (Lin and Tomlin, 2025), lacking real users’ attention lapses, linguistic noise, and irrationality (Takanobu et al., 2020). Most critically, both fixed datasets and non-adaptive simulators fail

to provide curriculum-based training that adjusts difficulty as the agent improves, limiting their effectiveness for complex multi-turn service dialogues requiring coherent long-term interactions.

Recent advances in self-evolution offer promising zero-data solutions for data-scarce business scenarios. These self-evolving methods (Chen et al., 2024; Zhao et al., 2025) employ self-play strategies where models generate both questions and corresponding answers, eliminating dependence on expensive expert-annotated data (Silver et al., 2017; Vinyals et al., 2019). However, applying these methods to service dialogues faces critical challenges. As shown in Figure 1, first, **unfair adversarial game**: in the service dialogue scenario, User Role-play Models can arbitrarily control outcomes, breaking the causal link between agent actions and task success. For example, simulators may reject agents regardless of response quality or accept based on turn count rather than persuasion effectiveness. Second, **real users are highly diverse**: without additional mechanisms, user behaviors easily fall into repetitive patterns.

To address these challenges, we propose SEAD (Self-Evolving Agent for Dialogue), the first self-evolving framework for multi-turn service dialogues. SEAD requires no large-scale annotated dialogue data, only user profiles and standard operating procedures as inputs. To avoid the unfair adversarial game where User Role-play Models arbitrarily control outcomes, we decompose the user side into two components: a profile generator that samples initial user states, and a role-play model that simulates responses. Crucially, only the profile generator participates in adversarial training by setting initial conditions. This design transforms participation into a betting game, where the user side must genuinely consider agent’s capability to identify the *golden training scenarios* where agents can succeed approximately half the time, enabling genuine adversarial learning. To maintain user diversity, the profile generator employs automated random sampling and consistency checks to ensure scenario diversity and authenticity.

We validate SEAD in outbound call services based on a real enterprise scenario. By enumerating 5 cooperation levels, 4 emotion levels, and 6 trust levels, we construct 120 initial user state combinations, evaluated through multi-level metrics. Experiments show SEAD significantly outperforms baselines using foundation models or large APIs. Small models trained via SEAD achieve su-

perior performance while drastically reducing costs. Notably, SEAD remains effective in data-scarce domains, enabling rapid deployment.

Our main contributions include:

- We propose the first self-evolving framework for multi-turn service dialogues that requires no large-scale annotated dialogue data.
- We design a decomposed user modeling mechanism that transforms participation into a betting game, forcing the user side to identify golden training scenarios and enabling genuine adversarial learning.
- We design a user scenario generation mechanism based on anonymized behavior patterns extracted from over 100k real dialogues, ensuring diversity, authenticity, and adaptive difficulty.
- Experiments show that SEAD achieves superior performance with significantly smaller model size: better at guiding users toward goals, more efficient in conversation flow, stronger in understanding user states, and more realistic in simulating authentic user behaviors, all without requiring large-scale annotated dialogue data.

## 2 Related Works

**Task-oriented Dialogue.** Task-oriented Dialogue systems are essential for managing complex inquiries in domains like e-commerce (Deng et al., 2024, 2025). While traditional neural models (Vinyals and Le, 2015; Wen et al., 2015; Shang et al., 2015; Li et al., 2016a) and early user simulations (Li et al., 2016b; Lewis et al., 2017; Wei et al., 2018) laid the groundwork, they face architectural limitations. Recent LLM-based approaches predominantly rely on static fine-tuning (Li et al., 2025; Ou et al., 2024; Zhu et al., 2025; Bernard and Balog, 2023), sometimes augmented by retrieval (Xu et al., 2024), multimodal inputs (Wang et al., 2025a; Gong et al., 2025), or reinforcement learning (Peiyuan et al., 2024). However, these methods often incur high annotation costs and lack real-world nuance. In contrast, SEAD introduces a fully dynamic user-agent interaction paradigm, bypassing data synthesis overhead to significantly enhance performance in complex scenarios.

**Self-evolving Agents.** Self-evolution leverages iterative generation and refinement with minimal supervision (Tesauro et al., 1995; Silver et al., 2017; FAIR et al., 2022). Early works utilized self-rewarding mechanisms (Chen et al., 2024; Yuan et al., 2024), evolving into "Coder-Tester" frameworks for verifiable domains like code (Lin et al., 2025; Wang et al., 2025b; Pourcel et al., 2025). Recent research has expanded this scope (Zhao et al., 2025; Huang et al., 2025; Sun et al., 2025), incorporating external environments and curated data to enhance evolution (Liu et al., 2025; Xia et al., 2025; Zhai et al., 2025). Distinctively, SEAD drives the self-evolution of a user role-play model and a customer service agent within a realistic environment, realizing genuine adversarial learning for complex multi-turn interactions.

### 3 Methodology

Service dialogue faces severe data scarcity, making self-evolving frameworks a promising solution. However, unlike objective tasks where correctness is verifiable, service dialogue outcomes are entirely subjective—users can arbitrarily control results regardless of agent quality, creating an unfair adversarial game. To resolve this, SEAD decouples user modeling into two components: a Profile Controller that samples initial states and participates in adversarial training, and a User Role-Play Model that focuses on realistic simulation without controlling outcomes. This design guides the Profile Controller to identify golden training scenarios (agent success rate  $\sim 50\%$ ) through initial state selection rather than mid-dialogue manipulation. Figure 3 illustrates the complete framework with four training phases. We first formalize the problem and define notations, then detail the framework components and training process.

#### 3.1 Problem Formulation and Framework Components

We aim to train a service agent that maximizes the reward  $R$ , measuring task completion and user satisfaction. Figure 2 illustrates our framework. First, the profile generator  $\pi_g$  samples initial user states  $p_0$  to create diverse user profiles, where  $p_\theta$  denotes the parameterized state distribution. Then, the user role-play model  $\pi_u$  enacts this user to interact with the service agent  $\pi_a$  through multi-turn dialogues. We model multi-turn service dialogue as a sequential decision process. At each turn  $t$ , the agent

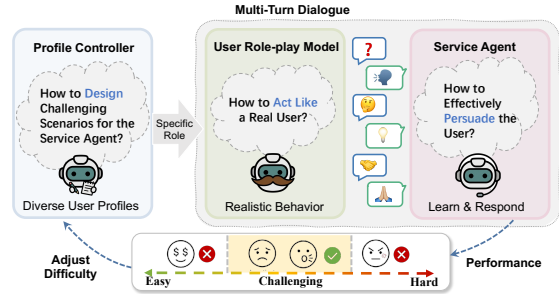


Figure 2: SEAD Framework Overview. SEAD consists of three components: (1) **Profile Generator** first creates diverse user profiles, then the (2) **User Role-Play Model** enacts these users to interact with the (3) **Service Agent**, training agents to adapt to any user. Finally, these dialogue data reflecting service agent capability returns to the Profile Controller, and initiates the next evolving loop.

observes dialogue history  $h_t = \{u_1, a_1, \dots, u_t\}$  and generates response  $a_t$ , where user response  $u_t$  and agent action  $a_t$  alternate. The agent maintains state estimate  $\hat{s}_t$  to guide action selection, maximizing cumulative reward  $R = \sum_{t=1}^T r_t$ . User state  $p_t = (c_t, e_t, tr_t)$  represents cooperation  $c_t$ , emotion  $e_t$ , and trust  $tr_t$ , which evolve based on agent behavior. User profile  $p_0 = (c_0, e_0, tr_0, \mathcal{B})$  defines initial state and behavior set  $\mathcal{B}$ , sampled from behavior library  $\mathcal{L}$ . Dialogue trajectory history is  $\mathcal{H} = (p_0, u_1, a_1, p_1, \dots, u_T, a_T, p_T, \text{outcome})$ . The role-play model generates  $u_t$  based on  $p_0$  and  $h_t$ , autonomously updating states. Its responses are determined by internal logic, ensuring outcomes depend on agent capability. Since the user side is naturally powerful, we only train the service agent to maintain  $\hat{s}_t$  and select  $a_t$ . Next, we detail the self-evolving training loop.

#### 3.2 Self-Evolving Training Loop

Figure 3 presents the complete self-evolving training loop through five interconnected phases alternating between online agent optimization and adaptive difficulty adjustment.

**Phase 1: Diverse User Profile Sampling.** The profile generator samples a batch of  $B$  initial profiles  $\{p_0^{(i)}\}_{i=1}^B$  based on dialogue trajectory history  $\mathcal{H}$ , which records completion rates for each state combination  $(c, e, tr)$ . In the first iteration, the generator performs random sampling to generate diverse profiles  $p_0 = (c_0, e_0, tr_0, \mathcal{B})$ , where  $\mathcal{B}$  is randomly sampled from user profiles in the library  $\mathcal{L}$ . In subsequent iterations, the generator employs statistics-driven sampling: prioritizing profiles with

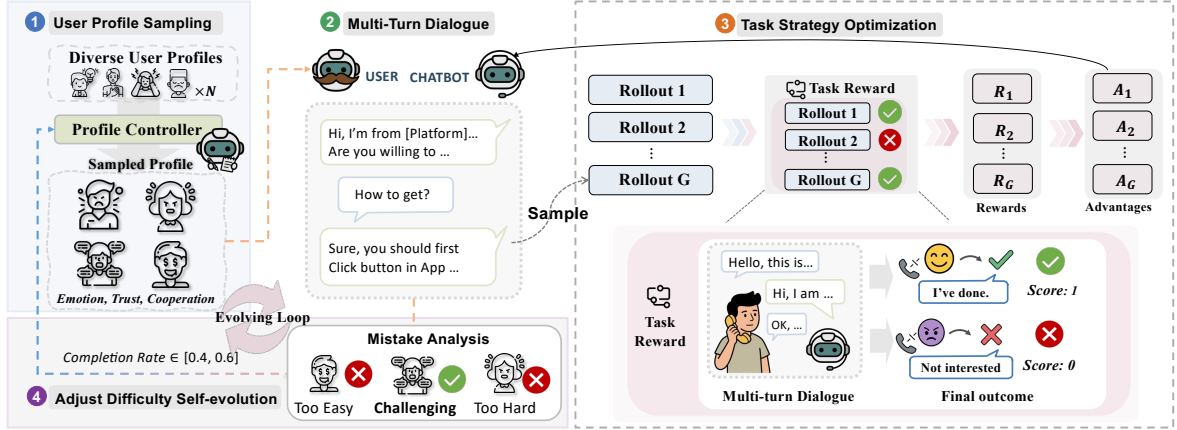


Figure 3: SEAD Co-evolutionary Training Loop. The controller samples initial states (Phase 1), which initialize dialogues producing trajectories (Phase 2), used to train the agent with rewards (Phase 3) and compute completion rates (Phase 4), which feed back to adjust sampling distributions, closing the co-evolutionary loop.

moderate difficulty where Completion Rates (CR) are close to 0.5, with sampling probability:

$$p_{\theta}(p_0|\mathcal{H}) \propto 1 - |CR - 0.5|. \quad (1)$$

The profile generator validates profile consistency and performs deduplication to prevent redundancy, ensuring diverse and high-quality profiles.

**Phase 2: Multi-Turn Dialogue.** Unlike prior static pre-collected dialogue full-filling approaches, SEAD enables dynamic multi-turn interactions between agents and simulated users. Each profile from Phase 1 starts a new conversation. The agent responds while keeping track of how the user feels, and the user reacts based on what the agent says and how the conversation is going. This back-and-forth creates diverse dialogues and shows how mistakes can pile up over multiple turns, just like in real customer service calls. We collect these complete conversations along with whether the task succeeded or failed.

**Phase 3: Task Strategy Optimization.** After each multi-turn dialogue in Phase 2, the collected trajectories are fed into online training. We optimize the Service Agent through conversation results, assigning rewards based on task completion status determined by the User Role-Play Model’s final state. Specifically, the task reward is defined as:

$$R_{task}^{(i)} = \mathbb{I}[\text{outcome}^{(i)} = \text{success}] \quad (2)$$

where  $\mathbb{I}[\cdot]$  is the indicator function, and  $\text{outcome}^{(i)}$  represents the dialogue outcome of trajectory  $\mathcal{H}^{(i)} = (p_0, u_1, a_1, p_1, \dots, u_T, a_T, p_T, \text{outcome})$ . The outcome is determined by the User Role-Play Model’s completion state at the final turn  $T$ : if the

user agrees to the service, outcome = success and  $R_{task} = 1$ ; otherwise,  $R_{task} = 0$ . Using Group Relative Policy Optimization (GRPO, elaborated in the following Section 3.3), we update agent parameters  $\theta_a$  to maximize expected rewards.

**Phase 4: Mistake Analysis and Self-Evolving Loop.** Unlike prior works that discard failed trajectories, we exploit them to guide evolution. We conduct statistical **Mistake Analysis** to identify where the model excels or struggles, categorizing configurations into *too easy* ( $CR > 0.6$ ), *too difficult* ( $CR < 0.4$ ), and *ideal* ( $CR \in [0.4, 0.6]$ ). This analysis feeds back to **Phase 1**, adjusting sampling distributions to maintain optimal learning difficulty near 50% completion rate. As the agent improves, the difficulty naturally escalates, creating a self-evolving curriculum that enables our 14B model to surpass 72B models and commercial APIs.

### 3.3 Training Optimization

Note that the User Role-play Model naturally dominates task-oriented dialogues and can arbitrarily determine outcomes. Without training, the User Role-play Model (URM) already exhibits realistic behaviors (Section 4.6). However, training URM degrades role-play quality, as URM will force success rates around 50% by directly accepting or rejecting regardless of agent performance, collapsing into extreme responses that ignore agent strategies. Therefore, we optimize only the service agent, which prevents the simulator from prioritizing adversarial outcomes over realistic role-play while reducing GPU memory by 50%. To train the Service Agent, we employ GRPO (Shao et al., 2024). This method eliminates the need for a separate value network,

significantly reducing computational resource requirements. Training computes advantages relative to the batch average:

$$A^{(i)} = R^{(i)} - \frac{1}{N} \sum_{j=1}^N R^{(j)} \quad (3)$$

where  $A^{(i)}$  is the advantage for trajectory  $i$ ,  $R^{(i)}$  is its reward, and  $N$  is the group size (number of sampled trajectories). The policy is then updated via gradient:

$$\nabla_{\theta_a} \mathcal{L} = \mathbb{E}_{\mathcal{H} \sim \pi_a} \left[ \sum_{t=1}^T A^{(i)} \nabla_{\theta_a} \log \pi_a(a_t | h_t) \right] \quad (4)$$

where  $\theta_a$  denotes agent parameters,  $\pi_a(a_t | h_t)$  is the policy distribution over action  $a_t$  given history  $h_t$  at step  $t$ , and  $\mathcal{H}$  represents trajectories sampled from the current policy.

### 3.4 User State Space Design

Our scenario is an outbound call service, where user profiles are extracted from real enterprise dialogues. We identify common behavior patterns such as questioning AI identity, expressing cost concerns, and showing attention lapses, which are randomly injected during training.

**Static Initial States.** Each user starts with an initial profile  $p_0$  containing three dimensions: cooperation  $c$  (willingness to cooperate), emotion  $e$  (emotional state), and trust  $tr$  (trust in the agent). By enumerating all combinations, we construct  $N = c_{\text{levels}} \times e_{\text{levels}} \times tr_{\text{levels}}$  initial states, covering a spectrum from highly resistant to fully cooperative users.

**Dynamic State Evolution.** User states evolve autonomously during conversations based on agent performance. The role-play model adjusts states dynamically: effective responses improve user states, while poor interactions degrade them. This evolution reflects the agent’s interaction quality and serves as the basis for reward computation, simulating natural user reactions and ensuring training realism.

## 4 Experiments

### 4.1 Task Setting

We evaluate SEAD in the outbound call service domain, specifically focusing on restaurant service promotion tasks. The goal is for agents to successfully convince restaurant owners to participate in

promotional activities. This task requires agents to handle diverse user reactions, build trust, address concerns, and maintain engagement across multiple turns. Critically, our setting requires no dialogue data. Training only needs: (1) Standard Operating Procedures (SOP) defining dialogue flow; (2) task objective description; (3) user profile. Agents autonomously explore optimal strategies through environment interaction, eliminating dependence on large-scale annotated data. This enables SEAD to rapidly deploy in data-scarce new domains and discover effective strategies beyond existing data distributions.

### 4.2 Implementation Details

All components use Qwen2.5-14B-Instruct. The code is implemented based on the VeRL framework with batch size  $B = 60$  and learning rate  $\alpha = 1 \times 10^{-6}$ . Dialogues terminate when users agree (success), refuse (failure), or reach maximum turns  $T_{\text{max}} = 15$ . The state space consists of cooperation  $c \in [0, 4]$  (5 levels), emotion  $e \in [0, 3]$  (4 levels), and trust  $tr \in [0, 5]$  (6 levels), yielding 120 initial states. Each state combination randomly samples at most  $N_{\text{max}} = 200$  behavior combinations to ensure diversity. All experiments run on 8 NVIDIA A100 80GB GPUs with decoupled architecture—profile controller, User Role-play Model, and service agent never occupy memory simultaneously, reducing peak memory requirements.

### 4.3 Evaluation Metrics

**Service Agent Metrics.** We evaluate agent performance using: **Completion Rate (CR)**, percentage of dialogues where users actually agreed; **Average Turns to Target (ATT)**, average dialogue length for successful cases (lower is better); **User Portrait Accuracy (UPA)**, accuracy of predicting user states, computed as  $\text{UPA} = 1 - \frac{1}{3} \left( \frac{\text{MAE}_c}{4} + \frac{\text{MAE}_e}{3} + \frac{\text{MAE}_{tr}}{5} \right)$  where  $\text{MAE}_c$ ,  $\text{MAE}_e$ ,  $\text{MAE}_{tr}$  measure errors for cooperation  $c \in [0, 4]$ , emotion  $e \in [0, 3]$ , and trust  $tr \in [0, 5]$ ; **Emotion/Trust/Cooperation Improvement (EI/TI/CI)**, average state changes from initial to final state; and **Total Cost**: Cumulative inference cost in CNY for 1000 multi-turn dialogue samples (API-based models only).

**User Role-play Model Metrics.** To validate that our simulator influences success rates based on agent quality, we establish a rubric mixing perfect human agent dialogues, SEAD (our trained agents), and low-quality agent dialogues. We eval-

Method	Params	CR (%)	ATT ↓	UPA	EI	TI	CI	Total Cost (CNY)
<i>Foundation Models</i>								
Qwen2.5-14B-Instruct	14B	38.7	10.5 <sup>±2.1</sup>	0.883 <sup>±0.085</sup>	0.34 <sup>±1.11</sup>	0.68 <sup>±1.53</sup>	0.63 <sup>±1.58</sup>	0.00
Qwen2.5-32B-Instruct	32B	38.3	9.9 <sup>±2.15</sup>	0.899 <sup>±0.068</sup>	-0.11 <sup>±0.54</sup>	0.76 <sup>±0.91</sup>	2.25 <sup>±1.15</sup>	0.00
Qwen2.5-72B-Instruct	72B	39.0	<b>9.6</b> <sup>±2.18</sup>	0.818 <sup>±0.144</sup>	<u>0.51</u> <sup>±1.32</sup>	1.06 <sup>±1.72</sup>	1.18 <sup>±1.59</sup>	0.00
<i>Large Model APIs</i>								
GPT-4o	–	<u>44.2</u>	10.8 <sup>±2.10</sup>	0.867 <sup>±0.117</sup>	0.04 <sup>±0.97</sup>	0.97 <sup>±1.29</sup>	1.34 <sup>±1.42</sup>	727.28
DeepSeek-Chat	671B	31.6	11.3 <sup>±2.10</sup>	0.863 <sup>±0.084</sup>	-0.20 <sup>±0.97</sup>	0.27 <sup>±1.24</sup>	0.76 <sup>±1.50</sup>	87.36
Qwen3-235B	235B	32.3	10.4 <sup>±2.50</sup>	0.765 <sup>±0.170</sup>	-0.24 <sup>±0.83</sup>	0.80 <sup>±1.14</sup>	1.54 <sup>±1.50</sup>	69.36
LongCat-Flash	560B	42.2	10.0 <sup>±2.31</sup>	<b>0.925</b> <sup>±0.079</sup>	0.28 <sup>±1.15</sup>	<u>1.33</u> <sup>±1.57</sup>	<b>1.56</b> <sup>±1.46</sup>	23.08
<b>SEAD (Ours)</b>	<b>14B</b>	<b>52.0</b>	<b>9.6</b> <sup>±2.09</sup>	<u>0.912</u> <sup>±0.071</sup>	<b>0.63</b> <sup>±1.12</sup>	<b>1.57</b> <sup>±1.51</sup>	<u>1.55</u> <sup>±1.39</sup>	<b>0.00</b>

Table 1: Main results comparison. Params: Model parameters (B=billion, "-" indicates undisclosed or not applicable). CR: Completion Rate. ATT: Average Turns to Target. UPA: User Portrait Accuracy. EI/TI/CI: Emotion/Trust/Cooperation Improvement. Total Cost: Total inference cost for 1000 multi-turn samples. Best results in bold. Standard deviations are shown as superscripts where available.

uate five dimensions using GPT-5.1 with few-shot human annotations: **Humanness / Emotion / Trust / Cooperation** (5=human-like, 0=robotic) and **Violation** (0=smooth, 5=severe). In real scenarios, most users exhibit minimal noise (score 1: hesitation, pauses) rather than severe violations; our simulator achieves 1.15, matching real behavior.

#### 4.4 Baselines

We compare SEAD against two categories of strong baselines: **Foundation Models**. We evaluate three sizes of Qwen2.5-Instruct models (14B, 32B, 72B parameters) using carefully designed prompts that include task descriptions and standard operating procedures. These models represent the zero-shot/few-shot capabilities of state-of-the-art open-source language models without task-specific training.

**Large Model APIs**. We compare against four commercial closed-source models: GPT-4o (Hurst et al., 2024), DeepSeek-Chat (Liu et al., 2024), Qwen3-235B-A22B (Team et al., 2024), and LongCat-Flash-Chat (Team et al., 2025). All API methods use carefully engineered prompts optimized for dialogue tasks.

We do not compare with Supervised Fine-Tuning (SFT) methods due to the lack of available data and the prohibitive cost of manual annotation. Generally, SFT methods are upper-bounded by data quality and exhibit poor generalization. Our approach eliminates this dependency and handles diverse scenarios effectively.

#### 4.5 Main Results

Table 1 presents the main experimental results. Our method achieves the highest service dialogue completion rate of 52.0% using only a 14B parameter model, outperforming the second-best baseline GPT-4o by 17.6% (52.0% vs. 44.2%) and improving over the pre-training 14B model by 34.4% (52.0% vs. 38.7%). SEAD also achieves the lowest Average Turns to Target (ATT) of 9.6, demonstrating superior dialogue efficiency in completing tasks more concisely.

For user state tracking metrics, SEAD outperforms most baselines and achieves competitive performance with LongCat-Flash, the dialogue-specific model with 40× more parameters and extensive pre-training on service dialogue scenarios. While LongCat-Flash obtains the highest User Portrait Accuracy (0.925 vs. 0.912), SEAD demonstrates comparable results across emotional improvement indicators. Specifically, SEAD achieves competitive scores on EI (0.63 vs. 0.28), TI (1.57 vs. 1.33), and CI (1.55 vs. 1.56), with SEAD actually leading on EI and showing near-identical performance on CI. This demonstrates that our self-evolution approach with adaptive curriculum learning enables a compact 14B model to match the user understanding capabilities of a 560B dialogue-specialized model, while requiring zero annotated dialogue data and maintaining superior task completion performance.

#### 4.6 User Role-Play Model Performance

Table 2 validates our user role-play model’s realism and diversity. Specifically, to ensure evaluation reliability, we extract anonymized behavior pat-

Metric	Mean	Std	Quality
Humanness	4.67/5	0.48	Near-perfect
Emotion	4.77/5	0.52	Highly human-like
Trust	4.81/5	0.45	Highly human-like
Cooperation	4.77/5	0.63	Highly human-like
Violation	1.15/5	0.90	Human-like Behaviour

Table 2: User Role-play Model quality. Higher humanness scores indicate more realistic simulation; lower violation scores indicate cleaner communication. All humanness metrics near-perfect ( $>4.5/5$ ).

terns from over 100k real enterprise dialogues. We then instruct GPT-5.1 to perform the assessment using annotated few-shot examples, contrasting high-scoring human instances against low-scoring failed model outputs. All humanness metrics exceed 4.5/5 with low standard deviations, demonstrating highly realistic and reliable behavior that mirrors real-world interactions. The violation score of 1.15/5 reflects authentic communication patterns with natural hesitation rather than artificial cleanliness or severe disruptions. The Profile Controller successfully generates diverse users from cooperative to skeptical, capturing heterogeneity essential for robust training. Crucially, consistent high scores across three agent quality tiers—perfect human agents, SEAD (trained agents, and low-quality agents) confirm our simulator adapts responsively to different strategies rather than following scripts, providing meaningful training signals.

#### 4.7 Ablation Study

To validate our core design choices, we conduct ablation studies on three components: (1) decomposed user modeling (keeping the User Role-play Model fixed), (2) Profile Sampling (PS) for intelligent initial state selection, and (3) Mistake Analysis (MA) for adaptive difficulty evolution. We compare four configurations:

**Configuration 1: w/o MA + w/o PS + Train URM.** This variant removes both Mistake Analysis (MA) and Profile Sampling (PS), allowing the User Role-play Model (URM) to autonomously select initial states. Critically, the URM is trained adversarially alongside the Service Agent, optimized based on dialogue outcomes. This represents traditional self-play where both sides evolve competitively. However, this violates our core principle: since the LLM-based user side can arbitrarily dominate dialogues (refusing cooperation, hanging up), training it adversarially creates an unfair game where success no longer depends on agent skill.

**Configuration 2: w/o MA + w/o PS.** This removes both adaptive mechanisms but keeps the URM fixed. Without the Profile Controller, the system lacks both structured initial states and closed-loop difficulty adjustment. This tests whether decomposed modeling alone (fixed URM) suffices without any adaptive control.

**Configuration 3: w/o MA.** This retains Profile Sampling (PS) but disables Mistake Analysis (MA)—the adaptive difficulty mechanism in Phase 4 (Figure 3). The Profile Controller samples from predefined user profiles but does not analyze training outcomes or adjust distributions based on agent capability. This tests whether random sampling from structured profiles suffices, or if closed-loop adaptation is essential for identifying golden training scenarios (50% success rate).

**Configuration 4: SEAD (Full).** Our complete framework integrates all three components: (1) fixed URM, (2) intelligent Profile Sampling (PS), and (3) adaptive Mistake Analysis (MA). As shown in Figure 3, Phase 4 analyzes completion rates and adjusts sampling distributions, forming a closed loop that ensures optimal training difficulty while maintaining user simulator authenticity.

Table 3 demonstrates that all three components are essential for SEAD’s effectiveness. The **Train URM** configuration suffers from catastrophic reward hacking: the simulator prioritizes adversarial scores over realism, collapsing into extreme responses (arbitrary acceptance or hang-ups). This results in degraded humanness (**URM-H**: 3.3) and poor task performance (**CR**: 35.2%). In contrast, **w/o MA** yields a 94.9% **UPA** improvement over **w/o MA + w/o PS** (0.877 vs 0.450), proving that structured Profile Sampling ensures behavioral diversity. Notably, **w/o MA + w/o PS** shows anomalously high trust improvements (**TI**: 1.87) despite low UPA, revealing a bias toward unrealistically easy scenarios. Ultimately, **SEAD (Full)** achieves optimal balance by identifying “golden” training scenarios (approx. 50% success). Overall, SEAD achieves a 47.7% relative **CR** improvement over adversarial training while maintaining peak simulator quality (**URM-H**: 4.7), demonstrating that decomposed modeling and adaptive evolution effectively prevent reward hacking.

#### 4.8 Case Study

As illustrated in Figure 4, the interactions reveal high-fidelity adversarial dynamics. The User Role-play Model generates distinct corner cases, rang-

Configuration	CR (%)	ATT	UPA	EI	TI	CI	URM-H
w/o MA + w/o PS + Train URM	35.2	11.8 $\pm$ 2.5	0.156 $\pm$ 0.120	-0.45 $\pm$ 0.95	0.32 $\pm$ 1.10	0.89 $\pm$ 1.25	3.3
w/o MA + w/o PS	47.7	9.8 $\pm$ 2.16	0.450 $\pm$ 0.000	0.89 $\pm$ 0.81	<b>1.87</b> $\pm$ 1.20	<b>1.67</b> $\pm$ 0.87	4.6
w/o MA	50.2	<b>9.6</b> $\pm$ 2.12	0.877 $\pm$ 0.088	0.69 $\pm$ 1.15	1.65 $\pm$ 1.51	1.56 $\pm$ 1.45	<b>4.7</b>
<b>SEAD (Ours)</b>	<b>52.0</b>	<b>9.6</b> $\pm$ 2.09	<b>0.912</b> $\pm$ 0.071	<b>0.63</b> $\pm$ 1.12	1.57 $\pm$ 1.51	1.55 $\pm$ 1.39	<b>4.7</b>

Table 3: Ablation study results. **MA**: Mistake Analysis and adaptive difficulty evolution. **PS**: Profile Sampling with intelligent initial state selection. **Train URM**: Training User Role-play Model in adversarial mode. **CR**: Completion Rate. **ATT**: Average Turns to Target. **UPA**: User Portrait Accuracy. **EI/TI/CI**: Emotion/Trust/Cooperation Improvement. **URM-H**: URM Humanness score (0-5 scale, higher indicates more realistic user behavior). Standard deviations are shown as superscripts where available. Results demonstrate that all three components are essential for optimal performance.

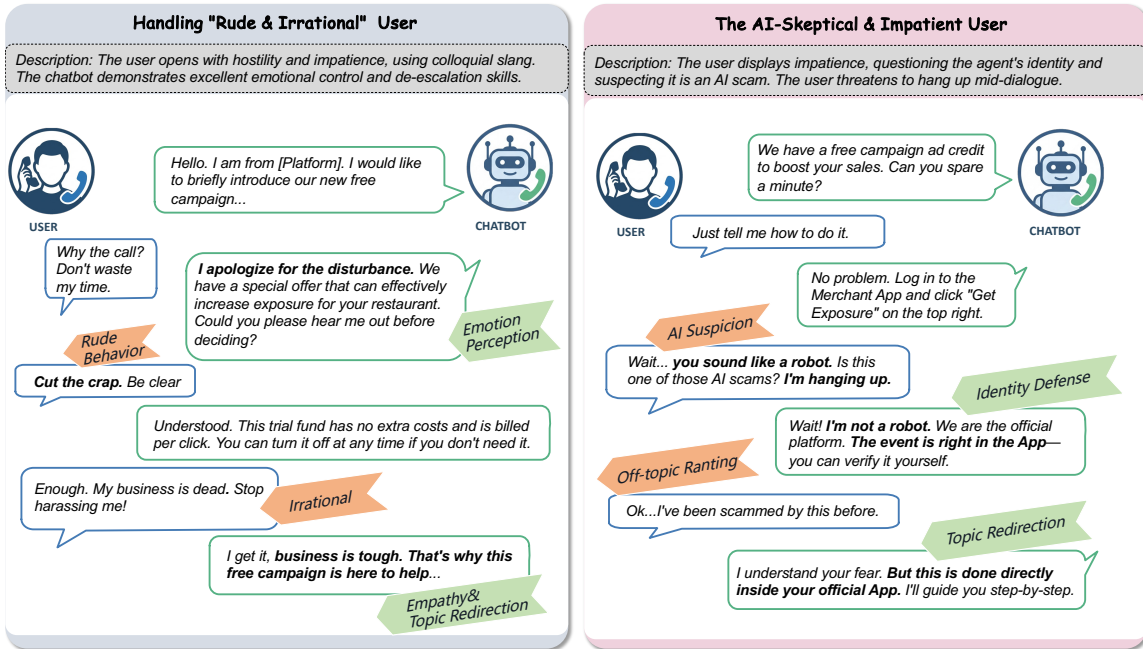


Figure 4: Case Studies of Challenging Interactions. The User Role-play Model generates heterogeneous personas via clustering, such as the "Rude & Irrational" user (Left) and the "AI-Skeptical" user (Right). The Service Agent demonstrates robustness learned from compound rewards, employing Empathy and Identity Defense strategies to prevent hang-ups and ensure task completion.

ing from irrational hostility (Left) to deep skepticism (Right). This heterogeneity stems from our user profile library derived from real-world enterprise data, which activates specific non-cooperative traits to create a rigorous training environment. In response, the Service Agent demonstrates exceptional adaptability, employing strategies like empathy and identity defense to retain users. This robustness stems from our reward mechanism and carefully designed user role-play model, which forces the model to prioritize task completion across diverse complex scenarios. By optimizing for state improvement, the agent learns to de-escalate conflicts efficiently, preventing premature hang-ups while avoiding the dialogue timeouts common in purely empathy-driven models.

## 5 Conclusion

In this paper, we presented SEAD (Self-Evolving Agent for Service Dialogue), a framework addressing data scarcity and user role-play fidelity in multi-turn service dialogues. By decoupling user modeling into a Profile Controller for curriculum learning and a User Role-play Model for authentic interaction, SEAD circumvents traditional adversarial training fairness. Experiments show SEAD outperforms both Open-source Foundation Models and Closed-source Commercial Models with minimal parameters and zero annotation. Future work will enhance emotional perception and extend to broader scenarios.

## Limitations

As an early exploration of a zero-data self-evolving service dialogue system, SEAD has limitations regarding evaluation metrics and scenario diversity. First, while we currently prioritize task completion, real-world applications demand high user satisfaction; thus, future work must better assess the agent’s ability to perceive emotion and maintain user comfort beyond mere intent fulfillment. Second, we have not yet extended our method to multi-scenario environments. Given its independence from curated data, our framework holds promise as a resource-efficient foundation model for diverse service dialogues, a potential we plan to validate in subsequent studies.

## LLM Usage Statement

In the preparation of this manuscript, a Large Language Model (LLM) was employed solely for language polishing. All academic content, interpretations, and responsibilities remain entirely with the authors.

## References

- Nolwenn Bernard and Krisztian Balog. 2023. Mshopdial: A multi-goal conversational dataset for e-commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2775–2785.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*, pages 6621–6642. PMLR.
- Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3):1–45.
- Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818.
- FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.
- Ning Gao, Wei Zhang, Yuqin Dai, Ling Shi, Ziyin Wang, Yujie Wang, Wei He, Jinpeng Wang, and Chaozheng Wang. 2026. Reinforcing real-world service agents: Balancing utility and cost in task-oriented dialogue. *arXiv preprint arXiv:2602.22697*.
- Ming Gong, Xucheng Huang, Chenghan Yang, Xianhan Peng, Haoxin Wang, Yang Liu, and Ling Jiang. 2025. Mindflow: Revolutionizing e-commerce customer support with multimodal llm agents. *arXiv preprint arXiv:2507.05330*.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.
- Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiabin Huang, Haitao Mi, and Dong Yu. 2025. R-zero: Self-evolving reasoning llm from zero data. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- 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*.
- Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016a. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1192–1202.
- Xiangci Li, Zhiyu Chen, Jason Ingyu Choi, Nikhita Vedula, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2025. Wizard of shopping: Target-oriented e-commerce dialogue generation with decision tree branching. *arXiv preprint arXiv:2502.00969*.
- Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016b. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*.
- Jessy Lin and Nick Tomlin. 2025. User simulators bridge rl with real-world interaction.
- Zi Lin, Sheng Shen, Jingbo Shang, Jason Weston, and Yixin Nie. 2025. Learning to solve and verify: A self-play framework for code and test generation. *arXiv preprint arXiv:2502.14948*.
- 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*.

- Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilya Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. 2025. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*.
- Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.
- Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170.
- Feng Peiyuan, Yichen He, Guanhua Huang, Yuan Lin, Hanchong Zhang, Yuchen Zhang, and Hang Li. 2024. Agile: A novel reinforcement learning framework of llm agents. *Advances in Neural Information Processing Systems*, 37:5244–5284.
- Julien Pourcel, Cédric Colas, and Pierre-Yves Oudeyer. 2025. Self-improving language models for evolutionary program synthesis: A case study on arc-agi. *arXiv preprint arXiv:2507.14172*.
- Ruifeng Qian, Shijie Li, Mengjiao Bao, Huan Chen, and Yu Che. 2022. Toward an optimal selection of dialogue strategies: A target-driven approach for intelligent outbound robots. *arXiv preprint arXiv:2206.10953*.
- Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khuu, 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*.
- Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Ling Shi, Yuqin Dai, Ziyin Wang, Ning Gao, Wei Zhang, Chaozheng Wang, Yujie Wang, Wei He, Jinpeng Wang, and Deiyi Xiong. 2026. Sage: A service agent graph-guided evaluation benchmark. *arXiv preprint arXiv:2604.09285*.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, and 1 others. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.
- Wangtao Sun, Xiang Cheng, Jialin Fan, Yao Xu, Xing Yu, Shizhu He, Jun Zhao, and Kang Liu. 2025. Towards agentic self-learning llms in search environment. *arXiv preprint arXiv:2510.14253*.
- Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation. *arXiv preprint arXiv:2005.07362*.
- Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, and 1 others. 2025. Longcat-flash technical report. *arXiv preprint arXiv:2509.01322*.
- Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).
- Gerald Tesauro and 1 others. 1995. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68.
- Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, and 1 others. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354.
- Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.
- Chaozheng Wang, Zongjie Li, Cuiyun Gao, Wenxuan Wang, Ting Peng, Hailiang Huang, Yuetang Deng, Shuai Wang, and Michael Lyu. 2026. Exploring multi-lingual bias of large code models in code generation. *ACM Trans. Softw. Eng. Methodol.*
- Haoxin Wang, Xianhan Peng, Huang Cheng, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Jiang Lin. 2025a. Ecom-bench: Can llm agent resolve real-world e-commerce customer support issues? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 276–284.
- Yinjie Wang, Ling Yang, Ye Tian, Ke Shen, and Mengdi Wang. 2025b. Co-evolving llm coder and unit tester via reinforcement learning. *arXiv preprint arXiv:2506.03136*.
- Wei Wei, Quoc V Le, Andrew M Dai, and Li-Jia Li. 2018. A goal-oriented neural conversation model by self-play.
- Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1711–1721.

- Peng Xia, Kaide Zeng, Jiaqi Liu, Can Qin, Fang Wu, Yiyang Zhou, Caiming Xiong, and Huaxiu Yao. 2025. Agent0: Unleashing self-evolving agents from zero data via tool-integrated reasoning. *arXiv preprint arXiv:2511.16043*.
- Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2905–2909.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *ICML*.
- Yunpeng Zhai, Shuchang Tao, Cheng Chen, Anni Zou, Ziqian Chen, Qingxu Fu, Shinji Mai, Li Yu, Jiayi Deng, Zouying Cao, and 1 others. 2025. Agentevolver: Towards efficient self-evolving agent system. *arXiv preprint arXiv:2511.10395*.
- Yulong Zhang, Tianyi Liang, Xinyue Huang, Erfei Cui, Guoqing Wang, Xu Guo, Chenhui Li, and Gongshen Liu. 2026. Consensus entropy: Harnessing multi-*vlm* agreement for self-verifying and self-improving ocr. *Preprint*, arXiv:2504.11101.
- Yulong Zhang, Li Wang, Wei Du, Peilin Li, Yuqin Dai Zhiyuan Zhao, Lingyong Fang, Ziniu Liu, Ru Zhang, Huijia Zhu, and Gongshen Liu. 2025. Ncv: A node-wise consistency verification approach for low-cost structured error localization in *llm* reasoning. *Preprint*, arXiv:2510.02816.
- Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*.
- Jie Zhu, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, Chi Zhang, and Fang Kong. 2025. Evaluating, synthesizing, and enhancing for customer support conversation. *arXiv preprint arXiv:2508.04423*.