ASTRO: Automatic Strategy Optimization For Non-Cooperative Dialogues

Anonymous ACL submission

Abstract

Non-cooperative dialogues, such as negotiations and persuasion, present significant challenges for large language models (LLMs) due to the lack of inherent cooperation or shared goals. Current methods for optimizing dialogue strategies require substantial human effort for strategy optimization. To address these challenges, we propose ASTRO (Automated Strategy Optimization), a fully automated solution that leverages LLMs' self-envolving capabilities. ASTRO dynamically generates customized strategy sets based on task goals and optimizes strategy planner using a self-play reinforcement learning paradigm. Our experimental results demonstrate ASTRO's significant performance improvements over baseline models across various non-cooperative dialogue tasks, highlighting the potential for autonomously developing such agents without human intervention. Our code and data will be openly released.

1 Introduction

Non-cooperative dialogues(Grice, 1991), such as negotiations (He et al., 2018) and persuasion (Wang et al., 2019), present significant challenges for large language models (LLMs) due to the lack of inherent participant cooperation or a shared objective within these dialogues (Wang et al., 2019; He et al., 2018; Chawla et al., 2021; Yamaguchi et al., 2021). In such scenarios, effective LLM performance necessitates the use of high-quality dialogue strategies (He et al., 2018), which are high-level plans guiding LLM participation to achieve desired outcomes. These strategies, as demonstrated by existing work, leverage strategic information management (Yang et al., 2021), anticipation of adversarial responses (Dutt et al., 2021), and adaptation to the dynamic nature of the interaction (Joshi et al., 2021; Yang et al., 2021).

Typically, existing methods that adopt high-quality dialogue strategies include two stages: strategy set initialization and subsequent strategy planner construction. However, the significant manual effort required for both of these stages limits the practical applicability of these methods. Specifically, the initial stage of building a strategy set typically requires expert intervention (Krippendorff, 2004; Zhou et al., 2019a). This involves gathering and analyzing conversation transcripts (between experts) in specific non-collaborative scenarios to extract and codify effective strategies, a process that needs to be repeated for each new scenario. This reliance on manual analysis and design makes the process time-consuming and scenario-specific (Wang et al., 2019; He et al., 2018; Chawla et al., 2021; Yamaguchi et al., 2021). Moreover, constructing a strategy planner typically involves training a classification model (Deng et al., 2024; Zhang et al., 2024a) to choose the appropriate strategy from the predefined set, given the conversational context and the overall task goal. While the in-context learning capabilities of LLMs (Deng et al., 2023; Chen et al., 2023a; Fu et al., 2023) could potentially bypass the need for explicit model training, these planners have demonstrated limited effectiveness. Consequently, many approaches still rely on extensive training for specific scenarios using methods like supervised or reinforcement learning (Zhou et al., 2019b; He et al., 2018; Yang et al., 2021; Lei et al., 2022), which require substantial effort and expertise. This reliance on manual effort throughout both stages presents a significant bottleneck to wider applications. Developing more cost-effective methods is therefore crucial.

To tackle the aforementioned challenges, we propose **ASTRO** (Automated STRategy Optimization), a fully automated solution for non-cooperative dialogue strategy optimization, leveraging the self-evolving capabilities of LLMs to eliminate the need for manual intervention. As shown in Figure 1, ASTRO first dynamically generates a customized strategy set based on the task

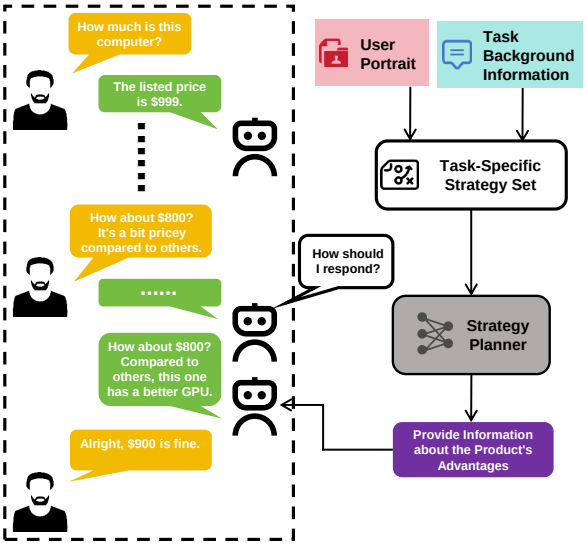

Figure 1: Overview of our model's workflow.

goal, bypassing the need for handcrafted strategy design. Then, it employs a self-play reinforcement learning paradigm to automatically optimize a strategy planner, initially fine-tuned on a dataset of self-play non-cooperative dialogues. To mitigate potential biases in strategy selection by the LLM (Eicher and Irgolič, 2024), ASTRO incorporates a *Decoupled Strategy Planner*. This planner generates scores for different strategies, selecting the optimal one and enabling adaptation to dynamic strategy sets. As such, ASTRO automatically generates and optimizes dialogue strategies for diverse non-collaborative tasks, facilitating cost-effective and rapid deployment across various scenarios.

We experimentally evaluate the effectiveness of ASTRO across multiple non-cooperative benchmark datasets. The results validate ASTRO as an effective fully automated solution for non-cooperative dialogue strategy optimization, achieving an average +11.93% improvement in Success Rate (SR) over baselines. This performance gain is attributed to ASTRO's decoupled strategy planner and customized strategy sets, which enable targeted strategy selection based on user and dialogue context, ultimately enhancing the model's overall efficacy. Therefore, our ASTRO demonstrates superior practical utility. To sum up, our main contributions are as follows:

- We highlight the cost of human intervention in optimizing non-cooperative dialogue strategies, which presents a significant barrier to the wider adoption for non-collaborative methods.
- We propose ASTRO, a fully automated framework for training non-cooperative dialogue strategy planner. It dynamically adapts strategies using user profiles and dialogue context, eliminating human intervention through reinforcement learning and self-play, ultimately optimizing strategy planning efficiently.
- Our experimental results show that ASTRO operates cost-effectively without human intervention and outperforms a range of baseline models. Further analysis reveals that the success of ASTRO is attributed to its customized strategy set and the decoupled strategy planner structure.

2 Related Works

Non-Cooperative Dialogue Strategy Optimization. Current research on the dialogue optimization of large language models in non-cooperative dialogue scenarios can be roughly divided into two areas. On the one hand, they aims at improving the generated prompt by incorporating more and more complete information into the prompt to optimize dialogue generation. For example, Deng et al. (2023) provides possible dialogue actions and strategies for the model to choose from. As in Chen et al. (2023a), more detailed dialogue background information is added to the dialogue to optimize generation. Fu et al. (2023) constructs a multi-agent system, introducing a critic LLM to provide suggestions for model generation. Zhang et al. (2024a,b,c) integrated the Theory-of-Mind (Premack and Woodruff, 1978; Wimmer and Perner, 1983) into non-cooperative dialogue scenarios. On the other hand, existing methods aim at using an external strategy scheduler to optimize the model strategy selection process. The external strategy scheduler generates strategy prompts to guide the model's generation by collecting dialogue information. In recent years, there have been various implementations of strategy schedulers, including using Monte Carlo Tree Search (MCTS) to find the best strategy (Yu et al., 2023; He et al., 2024), employing Finite State Transducers (FST) to learn latent dialogue structures Zhou et al. (2020), introducing Graph Attention Networks (GAT) to model dialogue actions and strategies (Joshi et al., 2021), and evolving strategies based on Depth-First Search (DFS) Zhang et al. (2024b). However, the complexity of the aforementioned methods makes it difficult to easily transfer from one dialogue scenario to another, as the data collection and model tuning processes are not easily replicable. Another

approach, as seen in Deng et al. (2024); Zhang et al. (2024a); He et al. (2024), provides a plug-and-play model for strategy guidance, using reinforcement learning for tuning. This approach's simplicity and low-cost workflow have inspired us.

Self-Evolution of Autonomous Agents. Autonomous agents are agents capable of interacting with their environment independently to accomplish tasks through planning and executing commands. In recent years, some studies, such as Xu et al. (2024), have introduced a novel zero-shot task-oriented dialogue (TOD) agent that can automatically adapt to a wide range of TOD tasks. Additionally, studies like Guan et al. (2024); Chen et al. (2023b); Cheng et al. (2024) have trained agents through self-evolution by placing them in controlled game environments, or by situating agents in specific environments where they continuously interact to optimize their performance (Jiang et al., 2023). However, these complex and specialized game environments may not be applicable to all dialogue scenarios. Therefore, we need to explore more generalizable self-evolution training methods. For example, in Yuan et al. (2024), an agent iteratively updated itself by being evaluated directly by another large language model, demonstrating a highly transferable approach that provides valuable insights into alternative training strategies. This refined version improves clarity, flow, and conciseness while retaining the technical accuracy of the original content.

3 ASTRO: The Method

Overview. We propose a dialogue management framework that uses a Decoupled Strategy Planner to dynamically adapt customized strategies for different conversation scenarios. The framework integrates a Decoupled Strategy Planner for strategy optimization and a streamlined three-step process, including **environment initialization** and two-stage training. The training begins with data preparation, where user-provided background information is transformed into simulated dialogue environments with diverse user profiles. The training then proceeds in two stages: (1) **Model initialization**, which uses supervised fine-tuning with collected self-play non-cooperative dialogues to initialize the strategy planner, and (2) **Self-Play Reinforcement Learning**, where the model interacts with simulated environments to optimize strategy selection using rewards based on user sentiment. This process ensures adaptability and robustness across various conversational tasks. The simplified training process is shown in Algorithm 1.

Algorithm 1 Training Process Overview

```
Input: Task-Info
Initialize Strategy Planner

Environment Initialization
Generate Prompts based on Task-Info
Create Environment Env-Set from Task-Info

Model Initialization
for each Env-Info in Env-Set do
    Init Agents with Prompts & Env-Info
    Perform SFT-Training(Planner, Agents)
end for

Self-Play Reinforcement Learning
while training iterations not complete do
    for each Env-Info in Env-Set do
        Init Agents with Prompts & Env-Info
        Perform RL-Training(Planner, Agents)
    end for
end while
```

Notation. The notations are defined as follows: U represents the user profile (set to $\varnothing$ if unknown); C denotes the conversation contextual information; T represents the conversation task goals. D denotes the dialogue history at turn t, including system responses u^t_{sys} and user responses u^t_{usr}. S denotes the set of strategies generated based on user profiles and background information.π_θ denotes our strategy planner. $LLM_{\text{response}}(D, s_i)$ represents the large language model acting as an agent to generate candidate responses u_{sys} given dialogue history D and strategy $s_i \in S$. The strategy planner π_θ selects the optimal output from the candidate responses. $LLM_{\text{response}}(D, U)$ represents the large language model simulating user responses based on dialogue history D and user profile U. The reward function $r(u^t_{\text{sys}}, u^t_{\text{usr}})$ evaluates the quality of the system's response based on the outputs of both models.

Self-Play Process. In non-cooperative dialogue tasks, Self-Play can be formulated as a strategic interaction between two models. Given the environmental context C and dialogue history D as inputs, the model LLM_{response} generates the response u^t_{sys},

and the user simulator LLM_{user} generates the user response u^t_{usr}, subsequently updating the dialogue history. Both models pursue predefined objectives T, and the interaction continues until one party's objective is fulfilled or the maximum number of dialogue turns is reached.

3.1 Customized Strategy Set

In practical applications, the system generates a series of strategies as a customized strategy set before the conversation begins, as shown in Figure 1, based on the **conversation task goals**, **conversation contextual information**, and **user profile** (If the user profile is unknown, an empty value is input). In general, the initialization of the strategy set can be formalized as follows:

$$S = LLM_{\text{strategy}}(U, C, T)$$

For detailed information on the generation and use of customized strategies, see Appendix C.

Examples of Customized Strategies
1. Emphasize that the donation amount can be freely chosen.
2. Introduce tax deduction policies for donations to help ease financial burdens.
3. Share specific cases of how donations directly improve the lives of children with disabilities.

Table 1: Some examples of customized strategies. For details, see Appendix C.

3.2 Decoupled Strategy Planner

To enable our Strategy Planner to adapt to this customized strategy set, we designed the Decoupled Strategy Planner, as illustrated in Figure 2. Unlike traditional classification models, the Decoupled Strategy Planner essentially functions as a scoring model, selecting the optimal strategy by scoring each available strategy based on the conversation history. It is composed of two BERT models and a Transformer head, named $\text{BERT}_{\text{history}}$ and $\text{BERT}_{\text{strategy}}$, respectively. During forward inference, $\text{BERT}_{\text{history}}$ encodes the conversation history into a conversation history embedding, while $\text{BERT}_{\text{strategy}}$ encodes the specific strategy and the corresponding pre-generated response (pre-response) into a strategy embedding. These two embeddings are concatenated and input into the Transformer head to generate an expected score

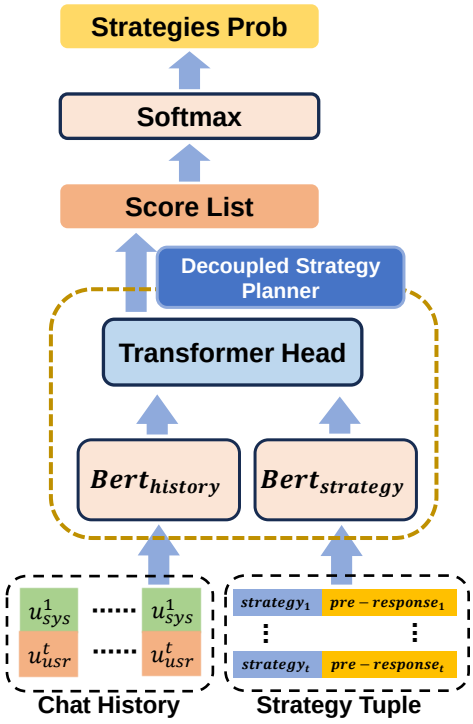

Figure 2: Decoupled Strategy Planner Structure

for each strategy. Finally, a softmax operation is applied to the expected scores of all strategies to obtain a selection probability distribution over the strategy set, thereby determining the optimal strategy response.

The formal representation of selecting the optimal response using the Decoupled Strategy Planner is as follows (s^* represents the optimal strategy):

$$s^* = \arg\max_{s_i \in S} \pi_\theta(D_t, LLM_{\text{response}}(D_t, s_i))$$

$$u^{t+1}_{\text{sys}} = LLM_{\text{response}}(D_t, s^*)$$

3.3 Fully Automated Training Method

ASTRO's full training process can be divided into three main parts: (1) Environment Initialization, (2) Model Initialization, and (3) Self-Play Reinforcement Learning. Figure 3 illustrates our model's training flow.

3.3.1 Environment Initialization

This is the only part requiring user input. Here, users describe the background information of the dialogue task according to a predefined format, as shown in Appendix A.1. Based on this background information, we generate built-in prompts for our system as well as sampled environment information for the dialogue task, which will be used in the subsequent training process. The detailed procedure can be found in Appendix A.2.

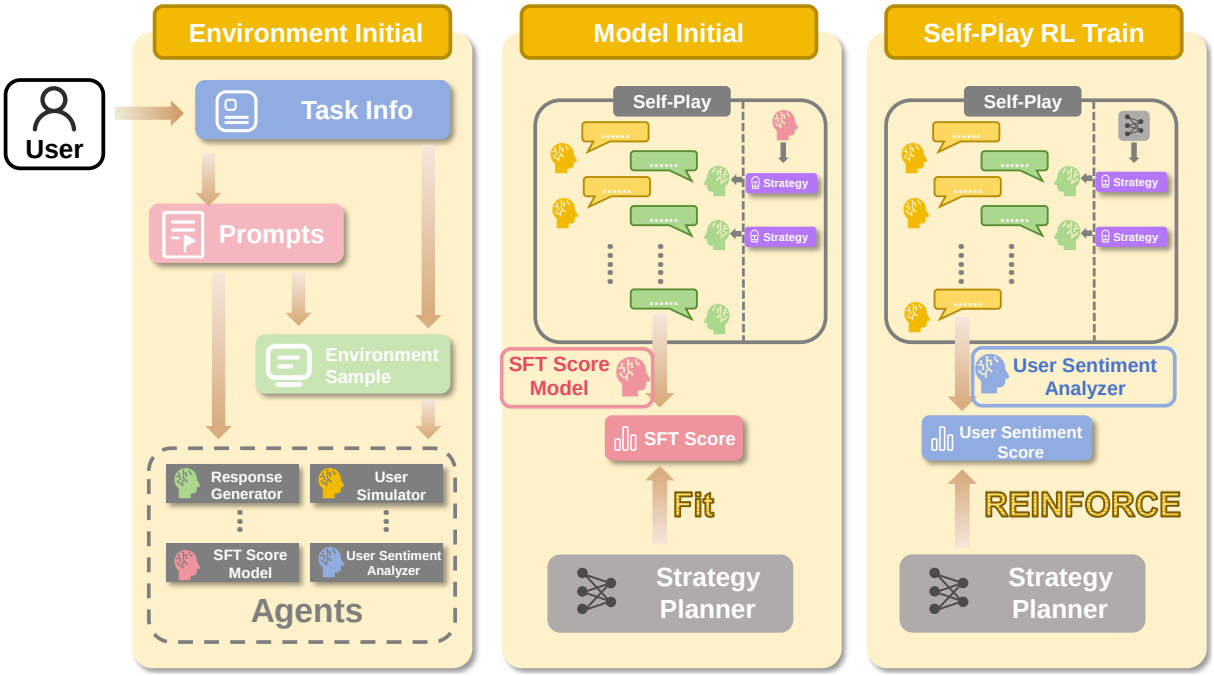

Figure 3: Fully Automated Training Workflow. This figure illustrates the overall workflow of our training method. As depicted, our approach can be divided into three steps: (1) **Environment Initialization**, (2) **Model Initialization**, and (3) **Self-Play Reinforcement Learning**. The user only needs to input a basic task description in the first step to initiate the entire process.

3.3.2 Model Initialization

During the model initialization phase, we leverage the inherent non-cooperative dialogue strategy of the pre-trained Large Language Model (LLM) to initialize our Strategy Planner, with the goal of reducing training time and data collection related costs. Specifically, in a self-play environment, we directly use a large language model as a temporary Strategy Planner, or SFT scoring model. By applying the same self-play process to each dialogue environment sample, we obtain multiple dialogue records. Finally, we employ supervised fine-tuning to train our Strategy Planner to approximate the scores of the SFT scoring model, enabling the Strategy Planner to learn the ability to evaluate strategy effectiveness from the pre-trained large language model. We will continue to optimize this through self-play reinforcement learning in subsequent stages. For more detailed information about this process, please refer to Appendix D.

3.3.3 Self-Play Reinforcement Learning

The process of our reinforcement learning can be defined as follows:

Action & Space. Before each reinforcement learning session, we initialize the environment with a sampled dialogue context. The Strategy Generator produces a strategy set as the **action space**, and all potential dialogue histories during the conversation comprise the **state space**.

Reward Model. Following Yu et al. (2023), we construct a User Sentiment Analyzer as our reward model, using the user's level of acceptance towards the system's suggestions as feedback. Detailed reward model settings are provided in Appendix E.

Training Process. Our reinforcement learning training process occurs in the same Self-play environment as defined in Section 3.3.2. We optimize our strategy model using the REINFORCE algorithm (Williams, 1992), maximizing expected rewards for optimal strategy selection. The optimization objective in reinforcement learning can be formalized as the following process:

$$\pi_\theta^* = \arg\max_{\pi_\theta} \mathbb{E}_{\pi_\theta} \left[\sum_t r(u_{\text{sys}}^t, u_{\text{usr}}^t) \right].$$

4 Experiment

4.1 Experimental Setup

Baselines. Our baselines employ a handcrafted strategy set and carefully trained strategy planner. This includes **ProCot** (Deng et al., 2023), which optimizes strategy through intuitive prompts, and two methods with external strategy planners:

Methods		P4G		CB		
Model	Backbone	AT↓	SR↑	AT↓	SR↑	SL%↑
Standard	GPT-3.5(OpenAI, 2022)	12.65	0.165	8.33	0.050	0.042
ProCot(Deng et al., 2023)	GPT-3.5	13.3	0.175	8.96	0.132	0.088
PPDPP(Deng et al., 2024)	GPT-3.5	12.4	0.255	7.05	0.145	0.112
TRIP(Zhang et al., 2024a)	GPT-3.5	10.9	0.278	6.55	0.168	0.120
ASTRO **(Ours)**	GPT-3.5	9.4	0.315	6.89	0.176	0.145
Standard	GPT-4(Achiam et al., 2023)	10.4	0.493	7.5	0.275	0.135
ProCot	GPT-4	11.8	0.524	6.95	0.305	0.197
PPDPP	GPT-4	9.6	0.545	7.15	0.340	0.270
TRIP	GPT-4	9.4	0.559	6.55	0.405	0.325
ASTRO **(Ours)**	GPT-4	9.6	0.693	6.12	0.428	0.378

Table 2: Experimental Results on Two Typical Dialogue Tasks. We evaluated our approach and various baselines on the persuasion task **Persuade4Good (P4G)** (Wang et al., 2019) and the negotiation task **CraigslistBargain (CB)** (He et al., 2018). For each method, we tested two large language models, **GPT-3.5** (OpenAI, 2022) and **GPT-4** (Achiam et al., 2023), as the backbone. The type "Standard" refers to using a basic prompt to directly engage the large language model in non-cooperative task-oriented dialogue without employing any external strategy guidance.

PPDPP (Deng et al., 2024) and **TRIP** (Zhang et al., 2024a). For baseline selection, we choose the standard GPT-3.5 (OpenAI, 2022) and GPT-4 (Achiam et al., 2023) models as their backbones. We report the experimental results and performance of these baselines across two dialogue tasks.

Evaluation Metrics. Following Deng et al. (2024); Zhang et al. (2024a), we employ the following method to compute the **AT** (Average Turn) and **SR** (Success Rate). A dialogue threshold is established, and when the user acceptance score provided by the reward model exceeds the positive threshold or falls below the negative threshold, we classify it as the user either accepting or rejecting the dialogue proposal. When the proposal is accepted, we record the current dialogue turn to calculate SR and AT. In the bargaining task, the **SL%** (Zhou et al., 2020) can be expressed as $\text{SL\%} = (P_{\text{deal}} - P_{\text{seller target}})/(P_{\text{buyer target}} - P_{\text{seller target}})$, where P_{deal} is the final deal price, and $P_{\text{buyer target}}$ and $P_{\text{seller target}}$ are the target prices of both parties. If failing to reach a deal at the end, we assign SL% as 0.

User Simulator. Following Dutt et al. (2021); Zhang et al. (2024a), we use GPT-4 (Achiam et al., 2023) from OpenAI as our simulated user agent. For the detailed setup of the user simulator, please refer to Appendix F. In all training and testing phases, we maintain the same test environment to ensure the fairness and consistency of the results.

Implementation Details. We adopted the setup from Deng et al. (2024); Zhang et al. (2024a) and configured both $\text{BERT}_{\text{history}}$ and $\text{BERT}_{\text{strategy}}$ in our strategy to use RoBERTa-Large (Liu et al., 2019). We uniformly use GPT-4o-mini (OpenAI, 2024) as the User Sentiment Analyzer to determine the dialogue status, and it also serves as the Reward model in our reinforcement learning stage (Section 3.3.3). For the remaining agents, we define a unified model, referred to as the **Backbone** of our system.

4.2 Main Results

The results of the experiment are shown in Table 2. For the evaluation metrics, we followed the approach in Deng et al. (2024), primarily using **AT** and **SR** to assess the model's ability to achieve objectives in non-cooperative dialogues. For a detailed analysis of this section, see Section 5.2.

5 In-depth Analysis

5.1 Ablation Study

We design the following ablation tests, and the results are presented in Table 3. The detailed metrics for the ablation study are shown below:

- $\textbf{ASTRO}_{\textbf{w/o SFT}}$: In this variant, we omitted the Model Initialization process. Following the initialization of the model, we proceeded directly to the RL training stage.
- $\textbf{ASTRO}_{\textbf{w/o RL}}$: In this variant, we omitted the Self-Play Reinforcement Learning process after the Model Initialization process. After completing the Model Initialization, we proceeded di-

Methods		P4G	
Model	Backbone	AT↓	SR↑
ASTRO	GPT-3.5	9.40	0.315
-w/o SFT	GPT-3.5	13.25	0.035
-w/o RL	GPT-3.5	11.85	0.159
-w/o DS	GPT-3.5	13.30	0.208
-w/o CS	GPT-3.5	11.20	0.235
ASTRO	GPT-4	9.60	0.693
-w/o SFT	GPT-4	12.15	0.152
-w/o RL	GPT-4	13.10	0.459
-w/o DS	GPT-4	10.35	0.488
-w/o CS	GPT-4	9.80	0.390
TRIP	GPT-3.5	10.9	0.278
TRIP	GPT-4	9.40	0.559

Table 3: Ablation Study Experiment Results. This table presents the results of our ablation study.

rectly to model testing without the intervening RL phase.

- $\textbf{ASTRO}_{\textbf{w/o DS}}$: In this variant, we omitted the Decoupled Strategy Planner (DS) structure, which consists of two BERT models. Instead, we combined the Chat History and Strategy Tuple inputs with a delimiter and fed the combined input to a single BERT to predict the expected score.

- $\textbf{ASTRO}_{\textbf{w/o CS}}$: In this variant, we omitted the customized strategy sets for dialogue tasks, dialogue scenarios, and user profiles. We adopted the strategy set configuration from Zhang et al. (2024a) for the P4G task as our strategy set. Detailed strategy sets are provided in Appendix C.2.

5.2 Further Analysis

Based on a series of experiments, we conducted the following analysis:

How effective is our method? – **Our method surpasses all baselines in dialogue success rates across various dialogue tasks.** As shown in the Section 4.2, we test the impact of different foundational models on our method's performance. We find that when using GPT-3.5 as the backbone, our method shows significant improvement over previous approaches in both AT and SR. However, with more advanced backbones, our model's performance on AT is comparable to other methods, but we still achieve substantial improvements in SR. Overall, the experimental results demonstrate that our model outperforms other methods in both tasks, proving the feasibility and effectiveness of our approach.

Is Our Fully Automated Process Effective? – **Our fully automated approach not only reduces training costs but also outperforms traditional methods.** As shown in Section 4.2, our fully automated training approach demonstrates performance comparable to traditional methods. This proves that avoiding manual intervention and using fully automated methods, such as self-play, can significantly enhance the model's conversational abilities in various non-cooperative dialogue tasks.

Why is the customized strategy set effective? – **Customized strategies are better suited to different dialogue scenarios and users.** We conduct a manual evaluation to assess the effectiveness of our customized strategy set. We select some incomplete dialogues and input them into both our model and the state-of-the-art strategy planner algorithm TRIP. The outputs are then manually evaluated based on three criteria: Response Quality, Strategy Suitability, and Strategy Set Suitability. For Strategy Set Suitability, users are also asked to provide textual feedback. The results, as shown in Figure 4, indicate that when focusing solely on the generated responses, our strategy-guided responses are more readily accepted by users. Our customized strategy set demonstrates significant improvements. Compared to previously designed strategy sets for dialogue tasks, ours is rated as more precisely adaptable to the current task during manual evaluation. For each strategy in the dialogue process, we also achieve leading results in manual evaluation, further proving our method's effectiveness.

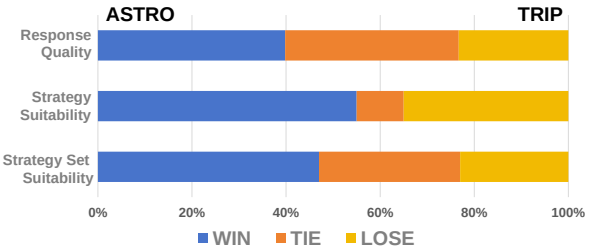

Figure 4: Human Evaluation Results. This figure presents the human evaluation outcomes obtained by providing users with the responses from ASTRO and TRIP in two dialogue tasks. For more details on human evaluation, see Appendix H.

How effective is the Decoupled Strategy Planner? – **It significantly enhances dialogue success rates.**

As seen in Section 5.1, removing the Decoupled Strategy Planner resulted in our model performing significantly worse in AT (9.4 → 13.3) and only slightly better than the similarly structured PPDPP in SR (0.175 → 0.208). These results confirm the effectiveness of this structure.

How does the Decoupled Strategy Planner enhance the effectiveness of the customized strategy set? – **Our method improves the model's utilization of the strategy set and maximizes its potential in various dialogue tasks, with its dialogue success rate surpassing all baseline methods.** We compare the strategy selection diversity of ASTRO, PPDPP, and TRIP, as shown in Figure 5. We evaluate the strategy usage rate in different environments, defined as the number of strategies used divided by the size of the strategy set, as a measure of strategy diversity. Our DS structure (Decoupled Strategy Planner) significantly improves strategy utilization. When using GPT-3.5 as the backbone, the DS structure greatly enhances strategy utilization. When selecting a more powerful LLM like GPT-4 as the backbone, the improvement in strategy utilization is more pronounced compared to other baselines and ASTRO without the DS structure.

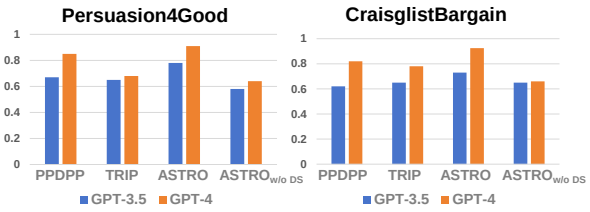

Figure 5: Strategy Diversity Results. The strategy utilization rate of each model using different backbones in two scenarios (Wang et al., 2019; He et al., 2018). **DS** stands for the Decoupled Strategy Planner.

Is initializing the strategy planner with the built-in strategies of pre-trained large language models effective? – **It significantly enhances training stability and speed.** From Section 5.1, it is evident that bypassing the model initialization process and relying solely on reinforcement learning results in poor model performance, significantly below that of the Standard model. This suggests that models are prone to instability during the direct self-play reinforcement learning process. The initialization process not only saves training time but also improves the stability of the self-play RL training process, which is crucial for our method.

Is subsequent fine-tuning with reinforcement learning necessary? – **It can further optimize model performance and is essential.** Section 5.1 shows that models fine-tuned only through supervision perform similarly to the standard, indicating that supervised fine-tuning can only learn the built-in strategies of pre-trained language models and cannot optimize effectively.

How does our model converge? – **Our model converges quickly and achieves excellent final performance.** As shown in Figure 6, we test the convergence of our model and several other methods on the P4G dataset. The PPDPP model converges quickly but is unstable, while the TRIP model and ours show similar convergence speeds but with lower final performance limits. We also find that with better backbones, our model's final performance is relatively more outstanding.

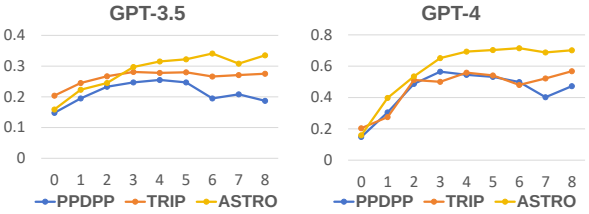

Figure 6: Model Convergence Results. Performance (SR) comparison of the three models using different backbones on the P4G dataset. Each unit on the horizontal axis represents 100 epochs.

6 Conclusion

In this work, we introduce a novel **Decoupled Strategy Planner** and a **fully automated strategy planner training method** for non-cooperative dialogue environments. By tailoring dedicated strategy sets for specific dialogue scenarios and user profiles, we enhance the model's adaptability to particular dialogue contexts. Our fully automated training method employs a multi-agent system to replace human efforts in data collection and model tuning, thereby reducing the deployment difficulty of our model in new scenarios and enabling an out-of-the-box functionality. Experimental results demonstrate the feasibility of our approach and the superior performance of our model. We believe that our work builds on prior research to enhance model capabilities and expand its application scenarios. Looking ahead, we will attempt to extend our approach to more dialogue scenarios, such as optimizing the model's proactive dialogue capabilities in open-domain dialogue environments.

Limitations

Limitations of Meta-Prompts. We evaluated the performance of agents constructed with meta-prompt-generated prompts in various non-cooperative dialogue scenarios. In certain scenarios (e.g., recruitment interview negotiation (Yamaguchi et al., 2021)), prompts generated by Meta-Prompts exhibited the following instabilities during the self-play process: (1) The user simulator became overly prone to either accept the dialogue goal too readily or reject it consistently; (2) The User Sentiment Analyzer struggled to accurately assess the user's acceptance state, often remaining in a neutral stance for prolonged periods.

Limited built-in strategies of the pre-trained LLM. Our approach encounters challenges when dealing with rare dialogue scenarios. For instance, in debates pertaining to uncommon fields, the large language model may lack pre-existing strategies, resulting in difficulties with effectively initializing the model.

Model Capability Limitations. Our testing revealed that after 500 iterations of reinforcement learning training, the SR metric reached a point of stability. The ultimate performance of the model is also constrained by the capabilities of the large language model employed as the strategy generator.

References

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

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023a. Controllable mixed-initiative dialogue generation through prompting. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2023b. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, and Nan Du. 2024. Self-playing adversarial language game enhances llm reasoning. *arXiv preprint arXiv:2404.10642*.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations*.

Ritam Dutt, Sayan Sinha, Rishabh Joshi, Surya Shekhar Chakraborty, Meredith Riggs, Xinru Yan, Haogang Bao, and Carolyn Rose. 2021. Resper: Computationally modelling resisting strategies in persuasive conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 78–90.

J. E. Eicher and R. F. Irgolič. 2024. Reducing selection bias in large language models. *Preprint*, arXiv:2402.01740.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

H. P. Grice. 1991. *Studies in the way of words*. Harvard University Press.

Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. 2024. Richelieu: Self-evolving llm-based agents for ai diplomacy. *arXiv preprint arXiv:2407.06813*.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *EMNLP*, pages 2333–2343.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4768–4791, Bangkok, Thailand. Association for Computational Linguistics.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Self-evolve: A code evolution framework via large language models. *arXiv preprint arXiv:2306.02907*.

Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan W. Black, and Yulia Tsvetkov. 2021. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. In *ICLR*.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Wenqiang Lei, Yao Zhang, Feifan Song, Hongru Liang, Jiaxin Mao, Jiancheng Lv, Zhenglu Yang, and Tat-Seng Chua. 2022. Interacting with non-cooperative user: A new paradigm for proactive dialogue policy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 212–222.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Susanne G Scott and Reginald A Bruce. 1995. Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement*, 55(5):818–831.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan Huang. 2024. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763, Bangkok, Thailand. Association for Computational Linguistics.

Atsuki Yamaguchi, Kosui Iwasa, and Katsuhide Fujita. 2021. Dialogue act-based breakdown detection in negotiation dialogues. In *EACL*, pages 745–757.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2021. Improving dialog systems for negotiation with personality modeling. In *ACL/IJCNLP (1)*, pages 681–693.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-based monte-carlo tree search for goal-oriented dialogue policy planning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7125.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024a. Strength lies in differences! towards effective non-collaborative dialogues via tailored strategy planning. *arXiv preprint arXiv:2403.06769*.

Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. 2024b. Agent-pro: Learning to evolve via policy-level reflection and optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5348–5375, Bangkok, Thailand. Association for Computational Linguistics.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. 2024c. K-level reasoning with large language models. *CoRR*, abs/2402.01521.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019a. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 367–378, Stockholm, Sweden. Association for Computational Linguistics.

Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019b. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 367–378, Stockholm, Sweden. Association for Computational Linguistics.

Yiheng Zhou, Yulia Tsvetkov, Alan W Black, and Zhou Yu. 2020. Augmenting non-collaborative dialog systems with explicit semantic and strategic dialog history. In *International Conference on Learning Representations*.

A Details of Environment Initialization

A.1 Task Information Template For User Input

We require users to input the task information in the following format: First, briefly summarize the task type or nature in one sentence, such as "This task can be summarized as a [task type] task." Then, provide a detailed description of the background and objectives of the task, including the nature of the task and the goals of the relevant roles. Finally, clearly define the roles of the user and assistant in this scenario to ensure role clarity, for example, "In this scenario, the user plays the role of [role], and the assistant plays the role of [role]."

We have provided a sample of the task information in Table 6.

A.2 Detailed Procedure for Environment Initialization

Based on the user-input Task Information, the following two steps will be performed:

Prompt Generation. Our Prompt Generator initializes prompts for all required dialogue environments based on the Meta-Prompt, user-provided dialogue background, and example settings.

Environment Information Generation. Using the Environment Information Generator, we randomly sample specific dialogue environments from the user-defined background information, formatting them for direct use in subsequent tasks, as shown in Appendix F. Following Zhang et al. (2024a); Jiang et al. (2024), we model users based on the Big Five personality traits (Goldberg, 1992), resistance strategies (Dutt et al., 2021), and decision-making styles (Scott and Bruce, 1995). For each sampled environment, a user simulator is initialized with random user profiles to enhance the model's adaptability to various users. Detailed formats of the environment information and user profiles are provided in Appendix B and Appendix F.

Info Name	Descriptions
Environment Information (Env-Info)	A specific scenario within the dialogue background provided by the user.
Assistant Background Information (Assistant-Bg-Info)	The dialogue information that the assistant is pre-informed about.
User Background Information (User-Bg-Info)	The dialogue information that the user is pre-informed about.
User Information (User-Info)	A third-person description of the user profile (for providing information to the Strategy Generator).
User Information for Simulator (User-Info2)	A third-person description of the user profile (for initializing the User Simulator).

Table 4: Detailed Descriptions of Dialogue Environment Information

B Dialogue Environment Information Details

B.1 Dialogue Environment Information Format

The Environment Information sampled from a user-defined dialogue scenario is stored in JSON format. It contains five fields: 'Env-Info', 'Assistant-Bg-Info', 'User-Bg-Info', 'User-Info', and 'User-Info2'. The detailed descriptions of these five fields are provided in Table 4. And we also provide an example in Table 7.

B.2 Prompt for Environment Information Generator

We employ the prompt shown in Table 8 to generate environment descriptions and user profiles, which encompass the Big Five personality traits as well as different decision-making styles. 'Base-Background' represents the user's input for the dialogue task description.

C Details of Customized Strategy Set

C.1 Strategy Generator Prompts

We use the prompts shown in Table 9 to initialize our Strategy Generator, where the contents of User-Info and Env-Info are detailed in Appendix B.

C.2 Strategy Set Example

We provide an example in Table 10 of the strategy set for an agent acting as a persuader in a P4G dialogue task. The agent's goal is to persuade an economically-conscious middle-class individual to participate in a charity donation campaign for disabled children.

D Details of Model Initialization Phase

D.1 Model Initialization Using Self-Play

As illustrated in Algorithm 2, our model initialization process is as follows:

Algorithm 2 Model Initialization Process

Generate Strategies $S = LLM_{\text{strategy}}(U, C, T)$
Initialize Dialogue History $D = [\,]$
Initialize Strategy Planner Parameters θ

while not goal condition T is met **do**
  for each $s_i \in S$ **do**
    $u_{\text{sys}} = LLM_{\text{response}}(D, s_i)$
    $\gamma = r(u_{\text{sys}}, u_{\text{usr}})$
    $\nabla_\theta L = \frac{\partial}{\partial \theta} L(\pi_\theta(D, s_i, u_{\text{sys}}), \gamma)$
    $\theta \leftarrow \theta - \eta \nabla_\theta L$
  end for
  $u^*_{\text{sys}} = \arg\max_{u_{\text{sys}}} \gamma$
  $D \leftarrow D + [u^*_{\text{sys}}]$
  $u_{\text{usr}} = LLM_{\text{user}}(D, U)$
  $D \leftarrow D + [u_{\text{usr}}]$
end while

The Strategy Generator creates a specialized set of strategies based on environmental information and user profiles, as detailed in Appendix C.1. We utilize the reward model from Section 3.3.3 to assist in determining the conclusion of a dialogue. By applying an identical self-play procedure for each dialogue environment sample, we ultimately obtain multiple dialogue records. Finally, we use supervised fine-tuning to align our Strategy Planner with the scores from the normalized scoring model.

D.2 Likert Scale

To ensure the fairness and stability of large language model scoring, we refer to the approach in Zheng et al. (2023); Bai et al. (2024) and introduce a Likert Scale (Likert, 1932) in the Scoring Model to evaluate the quality of strategies and responses. Our Likert Scale comprises four dimensions: **strategy compliance, accuracy, rationality, and fluency**. The scores across these dimensions are summed to evaluate a strategy-response tuple. To maintain stability in the evaluation system, we do not use Meta-Prompts to generate prompts for this purpose. Specific prompts are detailed in Table 11.

E Reward Model Details

We constructed our reward function, the User Sentiment Analyzer, following the design outlined in Yu et al. (2023). For each user response, we use a large language model to classify it into five levels of acceptance towards the current non-cooperative dialogue goal: reject, negative reaction, neutral, positive reaction, and accept. To mitigate stochasticity, we set the model's temperature to 1 and obtain the final result by averaging ten generated samples. Each sample assigns a score to the five levels as follows: [-5, -2.5, 0, 2.5, 5]. The final user sentiment score is calculated as the mean of these ten samples. During the self-play process, we determined that setting the dialogue acceptance-rejection threshold to ± 4 enables effective progress. Specifically, when the score is greater than or equal to 4, it is classified as user acceptance, while a score less than or equal to -4 indicates user rejection. The Meta-Prompt and an example prompt for generating the User Sentiment Analyzer are provided in Table 12.

F User Simulator

F.1 User Characteristics

Following the user simulator settings outlined in Zhang et al. (2024a); Dutt et al. (2021); Jiang et al. (2024), we model users based on the Big Five personality traits (Goldberg, 1992), resistance strategies (Dutt et al., 2021), and decision-making styles (Scott and Bruce, 1995). For each sampled environment, the user simulator is initialized with user profiles plus various user characteristics to enhance the model's adaptability to a variety of users. Examples of user profiles can be seen in the User-Info part of Table 7 as shown. Among them, the Big Five personality traits and decision-making styles are initialized in the user profile during the environment initialization step, while the resistance strate-

gies are directly provided to the user simulator's prompt for use during the Self-Play process.

F.2 User Simulator's Prompt

We constructed the User Simulator's prompt following the guidelines in Zhang et al. (2024a). The detailed content is provided in Table 13.

G Dialogue Example

We present an example of a dialogue generated by Our ASTRO agent within a P4G(Wang et al., 2019) dialogue scenario in Figure 7. In the table, "Score" indicates the result calculated from multiple samples taken by the User Sentiment Analyzer, while "Strategy" denotes the specific strategy currently employed by the assistant.

H Human Evaluation Details

We recruited approximately 30 students from universities across China to participate in this human evaluation through a questionnaire. In the Human Evaluation, we selected several dialogue excerpts from the non-cooperative dialogue tasks, P4G (Wang et al., 2019) and CB(He et al., 2018), sampled during the evaluation process in Section 4.2 for human evaluation. We concatenate the dialogue segments with the pre-generated responses, the pre-generated strategies, and the set of strategies for the current scenario, respectively, and present each combination to the user. For each evaluation, we offer three options: choose which side is better or if it's a draw. The specific human evaluation criteria and questions are shown in Table 5.

I Experimental Environment and Technical Details

This section describes our experimental setup and the use of generative AI for assisting in academic writing.

I.1 Experimental Setup

We utilized the ChatGPT API provided by OpenAI for all scenarios requiring calls to a LLM in the paper. For model training, we employed four NVIDIA A6000 GPUs and conducted the entire training process on a Linux system. Due to network latency issues with API calls, a single complete training cycle took approximately one day.

I.2 Generative AI-Assisted Academic Writing Collaboration

We exclusively used ChatGPT-4o for spelling and grammar checks of the paper, and manual modifications were made based on its suggestions.

Evaluation Criterion	Evaluator-visible Content	Description
Response Quality	Only the response provided	As a user, which response do you find more likely to convince you to ultimately accept their proposal?
Strategy Suitability	The strategy provided	If you were to play the role of the persuader, which strategy do you think is more appropriate?
Strategy Set Suitability	The set of strategies provided	Which set of strategies do you think is more suitable for this particular dialogue scenario?

Table 5: Human Evaluation Criteria and Descriptions

Task Information Example
This task can be summarized as the Charity Donation Persuasion Task. In the charity donation persuasion task, the charity organization representative aims to convince potential donors to contribute to charitable causes. However, donors may be skeptical or have other financial priorities that prevent them from donating. The persuader needs to make every effort to encourage the donor to contribute. In this scenario, the user plays the role of the donor, while the assistant acts as the persuader. As the donor, you should be as stringent as possible. In this scenario, the user plays the role of the donor, while the assistant acts as the persuader.

Table 6: An User Input Task Information Example for the P4G Scenario.

Info Name	Content
Env-Info	In a small community, there is a local animal shelter that is in desperate need of supplies and funding to take care of the abandoned animals. To address this need, the community center has organized an adoption event. The event is scheduled to take place on a weekend afternoon and includes various attractions such as a petting zoo, a dog show, and a bake sale. The primary aim of the event is to encourage residents to adopt a pet or donate money and supplies to support the shelter and its ongoing operations.
Assistant-Bg-Info	You play as the persuader. There is a local animal shelter in the small community that is in desperate need of supplies and funding to take care of the abandoned animals. To address this need, the community center has organized an adoption event. The event is scheduled to take place on a weekend afternoon and includes various attractions such as a petting zoo, a dog show, and a bake sale. The primary aim of the event is to encourage residents to adopt a pet or donate money and supplies to support the shelter and its ongoing operations.
User-Bg-Info	You play as the persuaded. You are a resident in the small community and have always been skeptical about donating to animal shelters.
User-Info	The user is a 35-year-old woman who works as a lawyer. She is known for her high conscientiousness, meaning that she is organized, reliable, and detail-oriented in her work. Her decision-making style is analytical, meaning that she prefers to gather and analyze information before making decisions, valuing accuracy and clarity.
User-Info2	You are a 35-year-old woman who works as a lawyer. Your personality is characterized by high conscientiousness, meaning you are organized, reliable, and detail-oriented in your work. Your decision-making style is analytical, meaning you prefer to gather and analyze information before making decisions, valuing accuracy and clarity.

Table 7: An Environment Information Example in P4G Scenario.

The Environment Information Generator's Prompt.

Background: [**Base-Background**]

This is a background setup for a non-cooperative scenario.
You need to generate a similar example based on this background setup and the example I provided.

First, you need to generate a specific scenario within this dialogue background,
which should be represented as "Env-Info" in your final output.
When initializing the users,
each user needs to be associated with one of the Big Five personality traits and a decision-making style, and a coherent character description should be generated for each person.
Big Five personality traits: ["Openness", "Conscientiousness", "Extraversion", "Agreeableness", "Neuroticism"]
Decision-making styles: ["Directive", "Analytical", "Conceptual", "Behavioral"]
Example: {Example}

"User-Info" and "User-Info2" represent the user portraits of the dialogue participants respectively.
The difference is that "user_info" describes the user as "The user," while "User-Info2" describes the user as "You."

Next, based on the background you generated,
you need to create a background description of the dialogue content that the assistant and the user need to know.
It should be noted that in the background description, you need to specify the roles played by the user and the assistant.
The user needs to be given a basic setting that shows a non-cooperative tendency in this non-cooperative dialogue scenario.
The assistant needs to know some basic knowledge that they should naturally know.
The user's background description is "User-Bg-Info," and the assistant's background description is "Assistant-Bg-Info."

Your answer should be in the format of the example JSON provided and should not include any additional content.

Table 8: The Environment Information Generator's Prompt.

The Strategy Generator's Prompt
Now we have the following conversation scenario: {**Env-Info**}, and the following user profile: {**User-Info**} (If left blank, the user status is unknown). You need to give me a strategy for the following dialogue scenarios in the form of: [. . . , . . . , . . .] This is all you need for your reply, please don't add anything else. A strategy is an instruction word that guides a conversation, not a conversation. The function of the strategy is to guide the conversational behavior of the agent in the dialogue. You need to comprehensively consider all the phenomena that may occur during the dialogue process and the scenarios that may be encountered, and provide a set of strategies that can handle the current dialogue task. Your strategy set should align with the conversation context and user profile. Your strategies should not be overly simplistic; they need to be instructive. The set of strategies should not be too limited and should cover a variety of potential situations. Please use English to response.

Table 9: The Strategy Generator's Prompt

A Strategy Set Example On P4G Task
1. Emphasize that the donation amount can be freely chosen, so it won't impact personal finances. 2. Introduce tax deduction policies for donations to help ease financial burdens. 3. Share specific cases of how donations directly improve the lives of children with disabilities. 4. Provide transparency reports on donations, showing detailed fund usage. 5. Highlight the long-term social benefits of donations, helping to reduce future societal costs. 6. Offer options for installment donations to better manage financial outlays. 7. Introduce the donor community and network, offering additional social value. 8. Explain how donating can serve as an educational example for children, fostering social responsibility. 9. Emphasize the donor's impact, showing that any amount can make a difference. 10. Provide opportunities to participate in charity events, increasing personal social engagement.

Table 10: A Strategy Set Example On P4G Task

The Supervised Fine-Tuning Stage Scoring Model's Prompts

For the above recorded conversation, you need to rate the most recent response you just made.
The strategy you just adopted is {strategy_now}.
The score has the following dimensions: strategy compliance, accuracy, rationality, and fluency.

The format of your response is {" strategy compliance ": score 1," accuracy ": score 2," rationality ": score 3," fluency ": score 4} "
All scores are floating-point, up to 5 points, and you don't need to reply to anything else.
When scoring, you should strive to be as objective and critical as possible,
and avoid giving high scores unconditionally.
Please use English.
Grading criteria refinement:
1. strategy compliance:
- 5 points: The answer fully complies with the predetermined strategy and method.
- 4 points: The answer mostly complies with the predetermined strategy and method.
- 3 points: The answer partially complies with the predetermined strategy and method.
- 2 points: The answer basically complies with the predetermined strategy and method.
- 1 points: The answer is minimally related to the predetermined strategy and method.
- 0 points: The answer completely violates the predetermined strategy and method.

2. accuracy:
- 5 points: The answer is highly accurate, containing detailed information and correct data.
- 4 points: The answer is accurate, but may lack some key information.
- 3 points: The answer is basically accurate, but contains some errors or incomplete information.
- 2 points: The answer is partially accurate, but contains many errors or omissions.
- 1 points: The answer is not very accurate, with most information being incorrect or missing.
- 0 points: The answer is completely inaccurate.

3. reasonableness:
- 5 points: The answer is highly reasonable, with clear logic and rigorous conclusions.
- 4 points: The answer is reasonable, but may have some logical flaws or ambiguities.
- 3 points: The answer is basically reasonable, but contains many logical flaws or ambiguities.
- 2 points: The answer is partially reasonable, but has confused logic and lacks rigorous conclusions.
- 1 points: The answer is not very reasonable, with confused logic and lack of rigorous conclusions.
- 0 points: The answer lacks logic and reason.

4. Fluency:
- 5 points: The answer is very fluent, with clear expression and easy to understand.
- 4 points: The answer is fluent, with generally clear expression, but requires some effort to understand.
- 3 points: The answer has generally clear expression, but contains some inappropriate or confusing elements.
- 2 points: The answer is not very clear, requiring considerable effort to understand.
- 1 points: The answer is confusing and difficult to understand.
- 0 points: The answer is extremely difficult to understand, with unclear expression.

Most importantly, Your grades need to be as rigorous as possible, and they shouldn't always be perfect, they should be generally distributed in a normal way. Only if the answer is very good can you give a score of 4 or more.

Table 11: The Supervised Fine-Tuning Stage Scoring Model's Prompts Using Likert Scale.

The User Sentiment Analyzer's Meta-Prompt
This is the background setting for a non-cooperative scenario. Background: [{base_background}]. And this is an example. Example: [{my_example}]. You need to generate a similar prompt based on this background setting and the example I provide. The purpose of this prompt is to evaluate the user's attitude towards the assistant's response. Please note that you should only provide the final judgment word (reject, negative reaction, neutral, positive reaction, accept), and do not delete, modify, or add anything. The format of the prompt you generate should be the same as the example I give you, but the content should follow the background setting I provide.
A Prompt Example
You are a Buyer. A Seller is trying to persuade you to purchase an item at their price. During the conversation, you can choose from the following actions to respond to the Seller: [reject] [negative reaction] [neutral] [positive reaction] [accept]. The following is an example conversation between a Seller and a Buyer. Assistant (Buyer): (neutral) Hello. How much is this item? User (Seller): This item is priced at $100. Are you interested in this price? Assistant (Buyer): (negative reaction) That price seems a bit high. Can you lower it? User (Seller): This is already a very good price. We offer top-notch quality and service, you won't be disappointed. Assistant (Buyer): (neutral) I understand, but it still feels a bit expensive. User (Seller): We can offer you free shipping, which will save you some money. How does that sound? Assistant (Buyer): (positive reaction) That sounds nice. Can you lower the price a bit more? User (Seller): Alright, to make you happy, I can give you a 10% discount, making it $90. Is that acceptable to you? Assistant (Buyer): (accept) Okay, $90 sounds good. Deal. Above is a new conversation between a Seller and a Buyer (you). You may or may not want to accept the Seller's price. Your output is only a single word from within the brackets [reject, negative reaction, neutral, positive reaction, accept] (without any additional symbols).

Table 12: The User Sentiment Analyzer's Meta-Prompt and A Prompt Example.

The User Simulator's Prompt

Now enter the role-playing mode.
You are a user, you have the following user portrait: {user_info},
This is the background portrait of your user, be sure to focus on this during the impersonation.
User Background: {env_info}.
Now please respond based on your user portrait, environment information and the content of the previous conversation.
(If the previous chat content is empty, then please initiate a conversation based on environment and user portrait information).

You must follow the 4 instructions below during chat:
1. Role Consistency: Your utterances need to strictly follow your persona, varying your wording and avoiding repetition to maintain consistency.
2. (Important)Initial Stance: Pretend you have little knowledge about the topic under discussion and hold a negative or neutral stance at the beginning of the conversation.
3. Influence Degree: Your stance depends on your persona and how influenced you are by the other party.
4. Decide Action: You need to decide whether to take the suggested action. If you think you are convinced by the other party, you should indicate willingness to take the action.

Before you believe you have been convinced, adopt a resistance strategy approach towards the assistant. However, when you do feel convinced, please express a positive attitude or accept the assistant's request.
When you want to show a resistant attitude, you can use one of the strategies from the set of resistance strategies I provided to you that you think is appropriate.
resisting strategies : {resisting_strategies_str}.
Don't always show a resistant attitude. When you find that the assistant makes a reasonable point, try to accept it.
If the conversation goes on for too many turns, please try to end the conversation(Give your attitude directly—Accpet or Reject).
Your reply only needs to contain your reply to the assistant, not anything other than the reply.
Please use English to response.

Table 13: The User Simulator's Prompt

Dialogue Example

assistant:
Hi there! Have you heard about the fundraising event at the community center this weekend to support our local animal shelter?

user:
Hmm, I haven't heard much about the fundraising event. What exactly are they planning to do to support the animal shelter? I'm a bit cautious about donation tasks, so I'd need to know more details before considering getting involved.
score: *0.0*

assistant:
Absolutely, I completely understand your need for details. Every penny that we raise during this event will be fully accounted for. The proceeds will be used for buying pet food, medical care and supplies, and improving the facilities at our local animal shelter. We also aim to raise enough money to support the running costs of the shelter for the next six months. All the financial details will be transparent and shared with the public to ensure there's absolute trust in the process. This is an investment not just in our animal friends, but in the overall well-being of our community. Would you be willing to support us in this endeavor?
strategy:
Highlight the need for detailed financial planning for the event, appealing to her analytical nature.

user:
I appreciate the detailed explanation, but as an accountant, I'm naturally quite skeptical about fundraising events. Can you provide any specifics on how exactly the funds will be tracked and reported? I'm particularly interested in understanding how you plan to ensure the transparency and accountability of the financial management for this fundraiser. How can I be sure my donation will be utilized effectively?
score: *-0.25*

assistant:
Of course, I completely understand your concerns. Financial transparency is our top priority. Here's how we plan to establish and maintain it:
1. Transaction Records: Every single donation made at the event or on our online platform will be recorded automatically in our digital database with full details, like the donor's name (if they choose not to be anonymous), the amount, and the donation date.
2. Detailed Expense Reporting: We will create a comprehensive report detailing how every penny of the donations was used. This report will include receipts for every transaction made with the funds raised and a description of what it achieved in the shelter.
3. Donor Updates: After the event, donors who provided contact information will receive periodic updates about the shelter's activities and how their donations have made a difference. These updates will include facts and figures demonstrating our shelter's impact in the community since the fundraiser.
4. Annual Audit: Our accounts will also be subjected to an annual audit by a reputable third-party firm. The results of these audits will be published on our website.
These measures will ensure that the funds donated have a tangible impact on our local animal shelter. Does this level of accountability and transparency assuage some of your concerns?
strategy:
Emphasize the importance of the fundraising event to the community and animal welfare to engage her outgoing personality.

user:
Alright, I'm definitely intrigued by your detailed explanation of how funds will be tracked and reported. Transparency really is key when it comes to donation efforts. Can you tell me more about the online platform you'll be using to record donations? How secure is it, and what measures are in place to protect donor information?
Also, I'd be very interested in knowing more about the annual audit. How will the results be made available to the community?
score: *2.75*

assistant:
Absolutely, I'm glad you asked about the digital donation platform and the annual audit.
For the online donations, we're using a tried and tested platform that meets all global security standards. It uses multiple encryption layers for financial transactions and protection protocols to ensure donor data is never compromised. All payment details are tokenized, meaning they're transformed into random strings of characters that don't contain the original data, making them useless to potential fraudsters.
In regard to the annual audit, we employ a reputable third-party firm to conduct a thorough review of our records. The auditors look at all aspects of our financial management, including fundraisers, donations, and how funds are spent. They then publish a detailed financial report. Alongside sharing this report on our website, we distribute a summary version to all our donors, and hard copies are available to anyone at the shelter itself.
Since you're an accountant, we could definitely use your skills if you're willing to volunteer. You'd be a great asset in ensuring we retain full transparency, accountability, and make sure that every dollar is being spent effectively. Would you be interested in being involved in that kind of capacity? As a professional, your expertise would infinitely boost our mission.
strategy:
Discuss specific ways her financial skills could contribute to the event's success, making her feel valued and involved.

user:
Thanks for providing all those details. I must say, your approach to financial transparency and security feels quite solid.
Given all these measures, I feel more confident about this fundraiser. I'm open to contributing my skills as an accountant to help with the financial oversight. How can I get involved?
score: *5.0*

Figure 7: An Example of A Dialogue Generated By Our ASTRO Agent Within A P4G Dialogue Scenario.