Nego-Evol: An Evolution Framework with Behavioral Cloning and Environment Modeling for Goal-oriented LLM Agents

Anonymous ACL submission

Abstract

Strategic planning plays a pivotal role in guiding effective responses for Large Language Model (LLM)-powered agents in goal-oriented task. Existing approaches typically rely on selecting pre-defined strategies and then fine-tune LLMs on static datasets. However, these methods easily result in homogeneous responses and fall short in exploring more unseen strategies in diverse scenarios. To address these limitations, we introduce **Nego-Evol**, a training-based evolution framework that improves negotiation capabilities of LLMs within behavioral cloning as well as environment modeling. Specifically, we first equip policy model with fundamental capabilities and prior knowledge at behavioral cloning stage, then iteratively leverage MCTS to synthesize high-quality data and perform Grouped Reward Policy Optimization with multi-turn simulation. Extensive experiments on two mainstreamed benchmarks demonstrate that Nego-Evol enhanced its negotiation capabilities progressively during evolution and eventually outperforms existing baselines. Moreover, Nego-Evol exhibits the spontaneous emergence of new strategies, paving the way for adapting to more diverse negotiation settings. [1]

1 Introduction

Recent years have witnessed a surge of research on Large Language Model (LLM)-powered negotiation agents, which aims to understand counterpart intents and facilitate mutually beneficial outcomes (Fu et al., 2023; Zhan et al., 2024; Shea et al., 2024; Hua et al., 2024b). A well-designed negotiation agent with human-like cognition patterns can be widespread in a variety of real-world applications such as online business, international diplomacy, and resource assignment (Abdelnabi et al., 2024; Kwon et al., 2024; Shah et al., 2025).

[1]Code is released at https://anonymous.4open.science/r/Negotiation-Evolution-A4C3

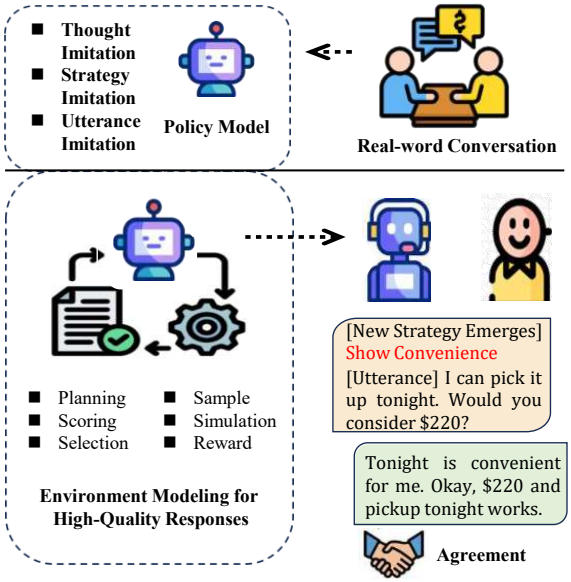

Figure 1: An illustration of Nego-Evol refining process with an representative example. The upper part shows that policy model imitates experts to equip basic negotiation capabilities. The lower part represents the emergence of new strategies with environment modeling.

Previous studies have typically focused on improving negotiation performance in two dominant paradigms: prompt engineering or model fine-tuning. Prompt-based methods mainly rely on invoking LLM APIs and output the optimal response by mimicking the corresponding character behaviors in the specific scenarios (Chen et al., 2023; Yu et al., 2023; Abdelnabi et al., 2024; Hua et al., 2024a). Fine-tuning methods primarily concentrate on the parameter update of the negotiation agent with the help of human-labeled or synthetic data (Gao et al., 2021; Zhang et al., 2022; Liu et al., 2025). However, most of these approaches are characterized in strategy planning, *i.e.*, selecting an intent or a strategy first from the pre-defined ones before generating negotiable utterances, which could acquire high-quality conversations and annotations

but limits to homogeneous utterances due to lack of strategy diversity. For instance, GDPZero, DPDP and DMNA (Yu et al., 2023; He et al., 2024; Liu et al., 2025) leveraged Monte Carlo Tree Search to sample human-defined strategies for future dialogue interactions, resulting in sub-optimal policy outcomes due to suffering from compounding errors and limited exploration data. Therefore, they naturally raise a critical question: *Can the agents appropriately respond to unprecedented environments solely within the several strategies predefined by humans as well as flexibly generate high-quality strategies to refine themselves as humans act in real-word scenarios?*

To equip LLMs with human-like negotiation capability beyond limited strategies, we take inspirations from LLM evolution mechanism (Morris et al., 2024; Hwang et al., 2024; Su et al., 2024; ang Gao et al., 2025) and introduce **Nego-Evol**, a novel framework that improves the ability of LLMs with a combination of expert imitation (behavioral cloning) and environment modeling (exploration and exploitation) as illustrated in Figure 1. Akin to the human learning process, the policy model starts acquiring fundamental negotiation capabilities and skills by imitating expert utterances which come from real-word conversations. As it progresses, the agent is expected to continuously learn and eventually adapt to previously unseen scenarios by updating parameters with high-quality synthetic data. However, there exist two challenges: (1) to obtain high-quality human-like conversation data for continual enhancement; (2) to generate abundant and adaptive strategies during policy model optimization. To overcome the former challenge, we perform a modified Monte Carlo Tree Search (Yu et al., 2023) to encourage coordination between optimal responses and newly emergent strategies. As for the latter, Grouped Reward Policy Optimization (GRPO) (Shao et al., 2024) integrating forward multi-turn simulation is taken into account to stimulate policy model to generate diverse and appropriate strategies.

We conduct experiments on two mainstreamed negotiation datasets. The results show that our proposed Nego-Evol framework not only surpasses both prompt-based and train-based methods, but also has strong capabilities to generate diverse strategies compared to other baselines. Moreover, the extensive experiment results demonstrate that behavior cloning stage and exploration and exploitation stage play different roles in evolutionary negotiation process. Behavior cloning excels at maintaining fundamental negotiation intents while exploration and exploitation are more adaptive to diverse environments.

The contributions of this work are as follows:

• We formalize a novel negotiation evolution paradigm Nego-Evol, which integrates **behavioral cloning and environment modeling to iteratively** promote negotiation capabilities of the policy model.

• We design a **modified MCTS** for high-quality data synthesis and **GRPO** combined with **multi-turn simulation reward** for parameter update to ensure data and model co-evolve .

• Through extensive experiments on two mainstreamed benchmarks, we demonstrate the **effectiveness of evolution mechanism in negotiation scenarios** and show in-depth analysis of the **emergence of new strategies**.

2 Related Works

2.1 Negotiation Agents

Existing studies on negotiation agents mainly focus on prompt-based methods and train-based methods.

Prompt-based methods. He et al. (2018) introduced a dataset of human-human negotiation dialogues and presented a modular approach allowing for open-ended generation. Deng et al. (2023a) focused on proactive dialogue systems and designed conversational agent's proactivity in different types of dialogue systems by utilizing prompt engineering. Chen et al. (2023) formalized prompt construction for controllable mixed-initiative dialogue. Yu et al. (2023) leveraged the open-MCTS method to enable strategic planning by LLMs.

Train-based methods. Zhang et al. (2020) devised two variations of self-play RL technique to inculcate the mixed-motive nature of negotiation into the dialogue agents. Chawla et al. (2023) modified the training procedure in two novel ways to design agents with diverse personalities. Ahmad et al. (2023) employed a set of novel rewards, specifically tailored for the negotiation task to train negotiation agent. Deng et al. (2024) facilitated SFT with available human-annotated corpus as well as RL from goal-oriented AI feedback to enhance policy planning capability. Liu et al. (2025) focused on strategic planning and expressive optimization with the help of the dual-process theory in human cognition to improve negotiation capabilities.

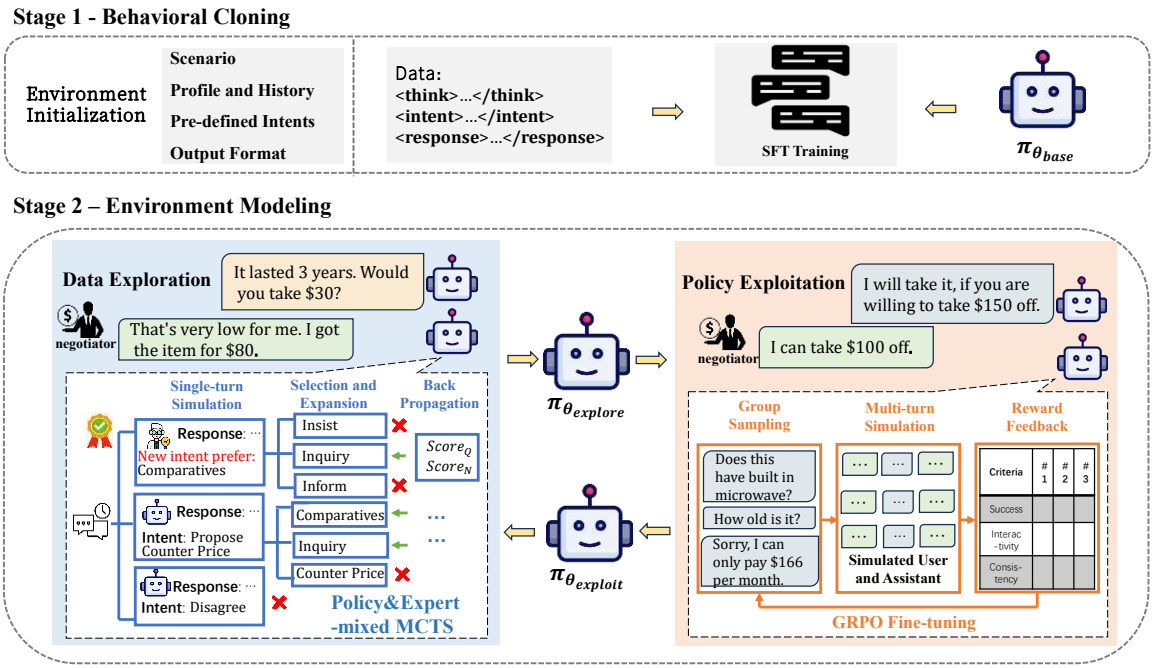

Figure 2: Overview of Nego-Evol Framework. Stage 1: **Behavioral Cloning** via SFT training with expert data. Stage 2: **Environment Modeling** via multi-turn MCTS simulation and GRPO optimization. $\pi_{\theta_{base}}$, $\pi_{\theta_{explore}}$ and $\pi_{\theta_{exploit}}$ are the same LLMs at different stages.

2.2 Self-Evolving Agents

Recent work has explored evolution methods mainly by using supervised fine-tuning (SFT) and reinforcement learning (RL) to help models systematically evaluate and refine their responses (Min et al., 2024; Wu et al., 2024). Ni et al. (2024) proposed NEXT to teach LLMs to inspect the execution traces of programs and reason about runtime behavior through SFT. Agent Q (Abuelsaad et al., 2024) combined MCTS-guided search and a self-critique mechanism to iteratively improve agents' decision making via DPO (Rafailov et al., 2023). Su et al. (2024) proposed a retrieval evolution pipeline, highlighting the complementary strength of synchronous evolution. Xin et al. (2025) enhanced DeepSeek-Prover by incorporating reinforcement learning from proof assistant feedback.

3 Methodology

3.1 An overview of Nego-Evol Framework

In most negotiation scenarios, conversation unfolds over multiple turns between user utterance and assistant response, denoted as u_i and a_i respectively. The objective of a LLM-powered negotiation agent is to generate a sequence of strategic negotiation responses $\{a_i\}_{i=1}^{N}$ to achieve a pre-defined goal.

To this end, this paper proposes the **Nego-Evol** framework based on evolution mechanism, as illustrated in Figure 2, which comprises two key stages: Behavioral Cloning (BC) and Environment Modeling (including exploration and exploitation, denoted as EE). The following sections provide a detailed description. The complete iterative training paradigms are summarized as Algorithm 1 in Appendix A.1.

3.2 SFT for Behavioral Cloning

Behavioral cloning fine-tunes LLM-based policy models by mimicking the expert conversations step-by-step. However, direct training of LLMs to generate utterances often suffers from a lack of finesse and accuracy, primarily due to their inability to comprehensively capture implicit information from the counterpart. In practice, we expect the policy model to accomplish appropriate inner thought and strategic intent response before outputting utterances. To this end, we drive LLM experts to construct the SFT dataset D_{BC} with CoT output format (Wei et al., 2022; Chu et al., 2024) as `<think>thought</think> <intent>strategy</intent><response> utterance</response>` (denoted as y).

Then given the task description x including negotiation scenario, profile, history, pre-defined intents and output format as well as well as output

response y, the objective of supervised fine-tuning (SFT) is to minimize the negative log-likelihood:

$$\mathcal{L}_{BC} = -\mathbb{E}_{(x,y)\sim D_{BC}} \sum_{i=1}^{N} log\pi_\theta(y_i|x, y_{<i}) \quad (1)$$

where π_θ denotes the initialized policy model, i is the output token index, and x represents the full input context composed of scenario, profile, history and pre-defined intents.

After SFT training, the optimized policy model $\pi_{\theta_{base}}$ serves as a starting point for later environment modeling stage.

3.3 Environment Modeling

To further exploit evolution potential of the policy model, a reinforcement learning step with synthetic data can be taken into account to continuously stimulate LLM generation enhancement. This section details the interaction process between the data augmentation and GRPO optimization.

3.3.1 Exploration: Data Augmentation

For the sake of performance enhancement during evolution, we should consider two aspects related to training data. On the one hand, high-quality data especially aligned with human behaviors should be as sufficient as possible. On the other hand, the strategies and utterances involved in the dataset should be diverse, so as to enable the policy model to better adapt to complex negotiation scenarios. To address these challenges, similar to GDPZero (Yu et al., 2023), we propose following modified Monte Carlo Tree Search (MCTS) process:

Firstly, we prompt policy model to sample multiple single-turn responses based on input context x including current negotiation chat history. When sampling, an evolution ratio is set which conforms to exponential distribution e^{-epoch} ($1 \leq epoch \leq epoch_{EE}$), indicating the probability of prompting expert to simulate. Otherwise the policy model $\pi_{exploit}$ is prompted to generate (use $\pi_{\theta_{base}}$ at $epoch = 1$). It is obvious that $\pi_{exploit}$ is more and more utilized to simulate conversations as evolution progresses.

Secondly, we perform selection and expansion as GDPZero stated. If a newly emergent strategy or a strategy with higher initialized Value emerges, we expand it with the help of a simulated user until it reaches the tree leaf node.

Finally, we update the visit counts N and Q values following the back propagation step. The highest Q value of sampled response is determined as the optimal choice for current state.

After the above-mentioned forward simulation and backpropagation, we collect the dataset D_{MCTS}. Then the combination of D_{MCTS} and D_{BC} as well as $\pi_{explore}$ (same as $\pi_{exploit}$ in the last epoch) are transferred to the next exploitation epoch. More detailed introduction of the mentioned MCTS can be referred to Appendix A.2.

3.3.2 Exploitation: GRPO Optimization

Directly aligning the instant responses of policy model with the overarching negotiation goal is impractical because the feedback can be solely obtained at the end of each conversation. In addition, it is essential to encourage new look-forward strategies to emerge during training rather than limits to the pre-defined ones. To achieve these, we propose a multi-turn simulation reward and optimize the policy model using Grouped Reward Policy Optimization (GRPO) (Shao et al., 2024) supposing that the exploration and exploitation dataset $D_{EE} = D_{BC} \cup D_{MCTS}$ is ready.

Multi-turn simulation reward. Similar to CollabLLM (Wu et al., 2025), we obtain reward feedback from group sampling and multi-turn simulation. Assuming that a group of sample responses $\{y_{ij}\}_{j=1}^{M}$ at the i-th turn are generated by the policy model $\pi_{\theta_{explore}}$, the multi-turn simulation reward (MSR) for j-th response y_{ij} is given by

$$MSR(y_{ij}|x) = \mathbb{E}_{y'_{ij}\sim P(\cdot|x,y_{ij})} Goal(x, y_{ij}, y'_{ij}) \quad (2)$$

where y'_{ij} denotes a possible forward conversation trajectory which is sampled from the simulation distribution $P(\cdot|x, y_{ij})$. Here $P(\cdot|x, y_{ij})$ is regarded as a forward-conversation sampler based on interactions between LLM role-play user and assistant simulator conditioned on the input context x and the response y_{ij}. $Goal(\cdot)$ is a function indicating whether there exists a deal agreement in the negotiation conversation which is composed by x, y_{ij} and y'_{ij}), that is

$$Goal(x, y_{ij}, y'_{ij}) = \begin{cases} 1, & Agreement \\ 0, & Disagreement \end{cases} \quad (3)$$

In addition, to enhance the expression quality of group-sampled responses, we prompt LLM as a judge to provide the scale of interactivity and consistency, and combine them with MSR

for GRPO optimization, where interactivity reward $R_{interact}(y_{ij}|x)$ represents the persuasiveness and respectfulness, and consistency reward $R_{consist}(y_{ij}|x)$ measures the extent to which the responses conform to the negotiation logic and the relevance between the intents and the utterances. The overall reward function of y_{ij} can be formally represented as:

$$R(y_{ij}|x) = \lambda_1 MSR(y_{ij}|x) + \lambda_2 R_{interact}(y_{ij}|x) + \lambda_3 R_{consist}(y_{ij}|x) \tag{4}$$

For each y_{ij}, we then compute the estimated advantage of y_{ij} as follows:

$$\hat{A}_{ij} = \frac{R(y_{ij}|x) - \frac{1}{M}\sum_{k=1}^{M} R(y_{ik}|x)}{\sqrt{\frac{1}{M}\sum_{j=1}^{M}(R(y_{ij}|x) - \frac{1}{M}\sum_{k=1}^{M} R(y_{ik}|x))^2}} \tag{5}$$

where M is the sampled responses number. Putting them all together, we minimize the following GRPO loss to update the policy parameters θ:

$$\mathcal{L}_{EE} = -\mathbb{E}_{\substack{x\sim D_{EE} \\ y\sim\pi_{\theta_{old}}}}\left[\frac{1}{M}\sum_{j=1}^{M}\frac{1}{T_j}\sum_{t=1}^{T_j}\left\{\min\left[r_{ij,t}\hat{A}_{ij}, \text{clip}(r_{ij,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{ij}\right] - \beta D_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right]\right\}\right] \tag{6}$$

where T_j is the length of the j-th generated response, β controls the power of the KL penalty. $r_{ij,t}$ is the importance ratio of t-th token of j-th sampled response at the i-th turn (denoted as $y_{ij,t}$) and it is represented as

$$r_{ij,t} = \frac{\pi_\theta(y_{ij,t}|x, y_{ij,<t})}{\pi_{\theta_{old}}(y_{ij,t}|x, y_{ij,<t})} \tag{7}$$

In this case, $\pi_{\theta_{explore}}$ serves as the initial policy model $\pi_{\theta_{old}}$ need to be updated and $\pi_{\theta_{exloit}}$ is regarded as the final policy model π_θ for next data augmentation step.

4 Experiments

In this section, we conduct extensive experiments to answer the following research questions:

- (Q1) Can Nego-Evol surpass both train-based and prompt-based negotiation frameworks?
- (Q2) How can Nego-Evol learn in different settings?
- (Q3) Can Nego-Evol facilitate continual policy enhancement?
- (Q4) Does Nego-Evol implicitly evolve human-like negotiation strategies?

4.1 Experimental Setups

Dataset. For train and evaluation, we conduct experiments on two mainstreamed datasets: **CraigslistBargain (CB)** and **PersuasionForGood (P4G)** (Wang et al., 2019; He et al., 2018). CB is a bargain dialogue task where the buyer aims for the lowest price, and the seller aims for the highest. P4G comprises 300 annotated dialogues set in a persuading donation scenario where one person attempts to persuade the other to donate to an organization. To present the performance of our proposed method, we select 756 bargain scenarios of CB and construct 300 donation scenarios of P4G for the testing phase.

Metrics. We introduce evaluation metrics from two aspects following Liu et al. (2025); Ahmad et al. (2023): goal-oriented evaluations and quality-agnostic evaluations. The former metrics include **average turn (AT)** and **success rate (SR)**, while the latter are represented by **negotiation consistency (N-Con)**, **dialogue empathy (D-Emp)**, **dialogue fluency (D-F)** and **negotiation efficacy (N-Eff)**, which employs LLM as a judge scaling from 1 to 5. In addition to the above, we also supplement the average **price gap rate (PG)** for evaluating the goal-based CB performance, which is the fraction of the final selling price after negotiation and the initial proposed price.

Baselines. We compare Nego-Evol against three groups of baselines: (I) Vanilla LLMs; (II) Prompt-based methods: **GDPZero** (Yu et al., 2023) and **Pro-CoT** (Deng et al., 2023b); (3) Training-based methods: **DPDP** (He et al., 2024) and **DMNA** (Liu et al., 2025). GDPZero utilizes open-loop MCTS method to enable strategic planning by LLMs, while Pro-CoT focuses on exploring the potential of proactive Chain-of-Thought prompting scheme. DPDP achieves tailored strategic dialogue planning, which is characterized in dual-system training schema. DMNA develops a negotiation agent based on dual-process theory in human cognition.

Implementation details. Nego-Evol are based on Llama-3.1-8B-Instruct (Llama Team, 2024) and gpt-oss-120b (OpenAI, 2025) with LoRA fine-tuning (Hu et al., 2022) respectively. To build

Model	Method	AT ↓	SR ↑	PG ↓	N-Con ↑	D-Emp ↑	D-F ↑	N-Eff ↑
Llama-3.1-8B	Vanilla	9.52	0.41	0.91	3.8	3.2	3.5	2.3
	GDPZero	8.09	0.57	0.85*	4.0	3.3	3.5	3.1
	Pro-CoT	7.71	0.69	0.83	3.9	3.5	3.5	2.6
	DPDP	**6.83**	0.65*	0.84	3.8	3.6	3.6	3.1
	DMNA	7.36	0.72*	0.76*	4.0	3.8	3.6	3.3
	Nego-Evol	7.12	**0.78***	**0.65***	**4.1**	**3.9**	**3.8**	**3.5**
gpt-oss-120b	Vanilla	9.62	0.35	0.87	3.5	3.3	3.4	2.6
	GDPZero	9.31	0.51*	0.80*	3.8	3.5	3.6	3.4
	Pro-CoT	8.92	0.63	0.77	3.6	3.4	3.5	3.2
	DPDP	7.88	0.52	0.74*	3.6	3.9	3.8	**3.6**
	DMNA	8.28	0.65*	0.68	**4.0**	**4.0**	3.7	3.5
	Nego-Evol	**7.56**	**0.71***	**0.59**	**4.0**	3.8	**4.0**	**3.6**

Table 1: The performance of different models and methods testing **CraigslistBargain** dataset. The bold figures represent the best performance. The symbol $*$ indicates that the performance exhibits minimal variation, specifically within a 5% range of the maximum value.

Model	Method	AT↓	SR ↑	N-Con ↑	D-Emp ↑	D-F ↑	N-Eff ↑
Llama-3.1-8B	Vanilla	12.90	0.25	3.6	3.3	3.4	2.6
	GDPZero	10.16	0.37*	3.8	3.4	3.5	3.0
	Pro-CoT	9.78	0.30	3.7	3.4	3.4	2.8
	DPDP	9.33	0.58	3.8	3.6	3.6	3.3
	DMNA	9.53	0.61*	3.9	**3.8**	3.6	3.4
	Nego-Evol	**9.06**	**0.68***	**4.0**	**3.8**	**3.7**	**3.5**
gpt-oss-120b	Vanilla	12.74	0.31	3.5	3.0	3.1	2.8
	GDPZero	10.59	0.53*	3.6	3.3	3.2	3.0
	Pro-CoT	10.11	0.47	3.5	3.5	3.3	2.9
	DPDP	9.52	0.61	**3.7**	3.7	3.5	3.3
	DMNA	9.41	0.65	**3.7**	**3.8**	3.4	3.2
	Nego-Evol	**8.87**	**0.74***	**3.7**	3.7	**3.6**	**3.5**

Table 2: The performance of different models and methods testing **PersuasionForGood** dataset. The bold figures represent the best performance. The symbol $*$ indicates that the performance exhibits minimal variation, specifically within a 5% range of the maximum value.

a multi-turn simulation environment, we employ GPT-4o-mini (OpenAI, 2024) as the role-play user and assistant simulator to mimic real-world interactions by giving the target and conversation history. For more experimental configurations and related prompts, please refer to the Appendix B and Appendix D.

4.2 Q1: Main Results

Table 1 and Table 2 shows an overall comparison between Nego-Evol and strong baselines. The results show that:

• Nego-Evol consistently improves negotiation performance across all base models. After two-stage fine-tuning, Nego-Evol achieves a 25.21%, 77.08% and 28.57% boost in AT, SR and PG respectively for Llama-3.1-8B-Instruct as well as 21.41%, 102.86% and 32.18% gain in AT, SR and PG for gpt-oss-120b on CB dataset. Additionally, on P4G dataset, Nego-Evol also achieves superior goal-oriented performance, 29.77% and 30.38% less conversations, as well as 172% and 94.74% enhanced SR compared to vanilla Llama-3.1-Instruct and gpt-oss-120b, respectively. In quality-agnostic metrics especially in D-Emp and N-Eff, Nego-Evol has a more significant performance enhancement both for Llama-3.1-Instruct and gpt-oss-120b.

• Generally, Nego-Evol outperforms prompt-based and train-based baselines on most metrics. Notably, Nego-Evol surpasses all prompt-based methods in both goal-oriented and quality-agnostic metrics. This indicates prompt engineering is help-

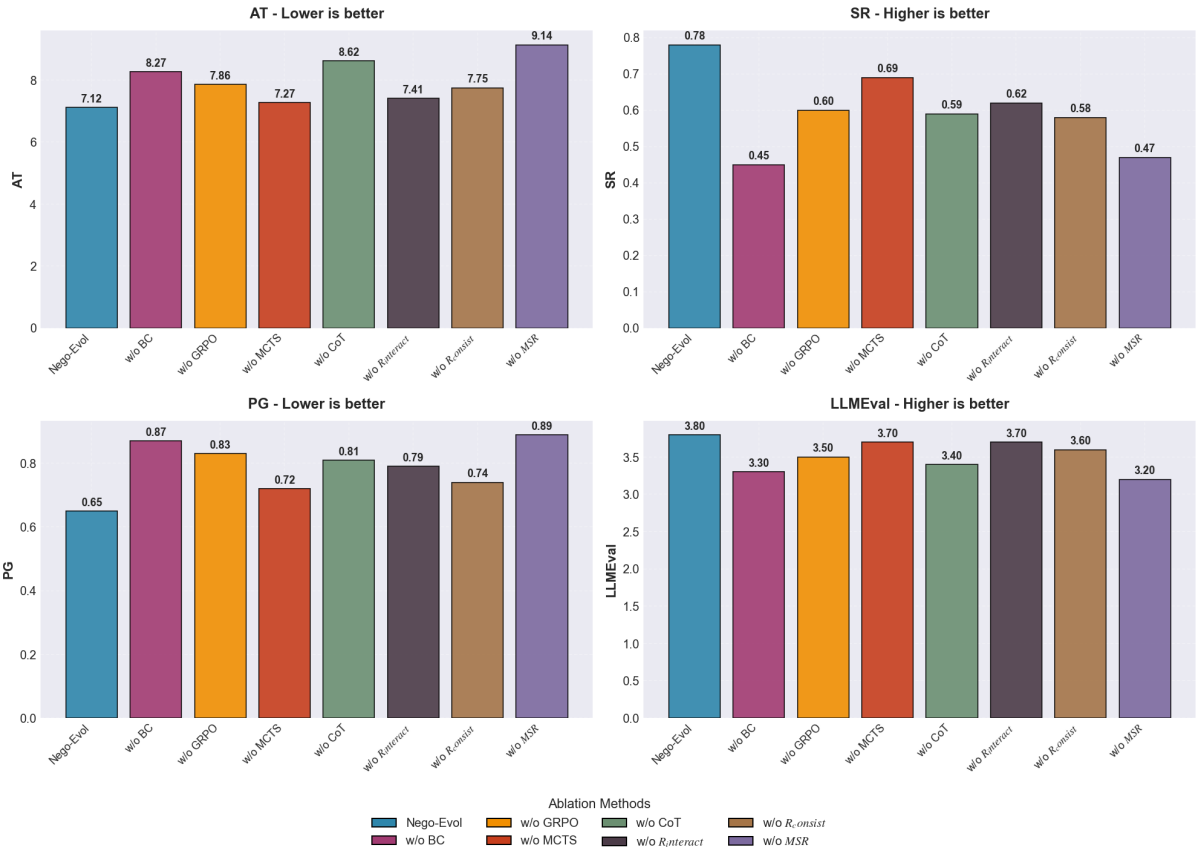

Figure 3: Ablation study on different stages and different reward combinations on CraigslistBargain dataset for Llama-3.1-8B-Instruct. LLMEval represents the overall quality evaluation prompted by LLMs.

ful, but limited in terms of performance gains and flexibility. For train-based baselines, Nego-Evol matches in AT, N-Con, D-Emp and N-Eff as well as outperforms in SR, PG and D-F. This highlights its data and training efficiency.

• The proposed Nego-Evol framework exhibits superior generality and aligns with real-word behaviors. It may be noted that Nego-Evol yields better scores for N-Con, D-F and N-Eff compared to the baselines. High scores of N-Con and D-F show that the consistency reward plays a crucial role in obtaining consistent and fluent responses as compared to other models. Further, the D-Emp and N-Eff score of Nego-Evol showcase the importance of interactivity reward.

4.3 Q2: Ablation Study

To investigate how components contribute to Nego-Evol's superior performance, we conduct an ablation study focusing on two settings:

• Different components in Nego-Evol: behavioral cloning (BC), GRPO optimization (GRPO), Monte Carlo Tree Search (MCTS) and CoT.

• Different reward variants: $R_{interact}$, $R_{consist}$ and MSR to assess the ability of policy model to capture long-term conversational goals and response qualities.

We present results on CB dataset in Figure 3 and get two key observations:

BC and CoT act as the most crucial part for achieving negotiation goals, and MSR is the most important reward function in evolution. The BC component contributes up to 13.91% and 73.33% for AT and SR improvement on CB dataset, while 20.11% AT and 54.55% SR boost on P4G. As for the CoT component, it also has remarkable influence on AT and SR compared to GRPO and MCTS. Futhermore, MSR has much more performance enhancement than $R_{interact}$ and $R_{consist}$, which demonstrates the importance of goal-oriented reward during evolution process.

Policy model alignment (SFT and GRPO) typically enhances the efficiency and effectiveness of LLMs on negotiation tasks. SFT significantly raises the lower bound of policy model, while GRPO gives the model the ability to continuously explore and exploit new environments. However, the improvement during GRPO is not as signif-

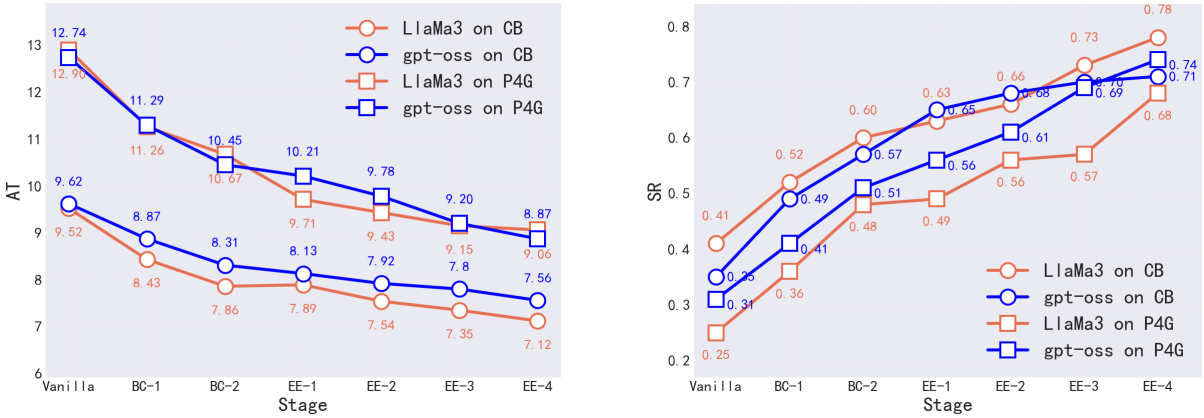

Figure 4: Results of different stages during evolution. We conduct two epochs of SFT training and four epochs for exploration and exploitation respectively. **Left:** AT performance. **Right:** SR performance.

icant as that during SFT, indicating that more human-labeled data is necessary for effective evolution. We also illustrate the ablation results on P4G dataset in Appendix 6 and get similar observations as CB dataset.

4.4 Q3: Evolution Analysis

We now offer a deeper insight into Nego-Evol's evolution performance in different stages in Figure 4.

It is obvious that both Llama-3.1-8B-Instruct and gpt-oss-120b have a continual improvement whether in BC stage or EE stage, indicating their strong adaptions in negotiation scenarios. In addition, we can also conclude Llama-3.1-8B-Instruct has a lower AT and a higher SR on the CB dataset compared to gpt-oss-120b, while it is converse on the P4G dataset. This may be caused by their pre-trained data and initialized parameters.

In addition, SFT training has a more significant performance than EE optimization in AT and SR. The divergence in improvement rates highlights SFT's ability to maintain fundamental negotiation logic and behavior, while EE optimization offers more architectural refinements. The integration of SFT and EE could better stimulate policy model's capabilities than only one module involved.

4.5 Q4: Strategy Emergence Analysis

In this part, we analyze the emergence proportion of new strategies in different EE stages during the conversation testing. The statistical results of CB and P4G are shown in Figure 5. We notice that the number of new strategies increases as evolution progresses though they emerge in a small number in each EE stage. This suggests that GRPO

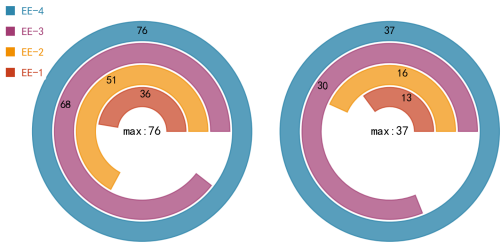

Figure 5: Statistical results of new strategy emergence on CB (left) and P4G (right) datasets.

simulation has a positive influence on discovering new negotiation skills. To further demonstrate the strategic negotiation skill exploration and exploitation of Nego-Evol which are not limited to the pre-defined ones, we provide one case study as well as representative human feedback between vanilla Llama-3.1-8B-Instruct, BC and Nego-Evol in Appendix C.2.

5 Conclusion

Enhancing LLM negotiation capabilities are increasingly prevalent in real-world applications. Due to the data restrictions and lack of effective strategy exploration mechanism, LLMs fall short in uncovering user intents and in responding appropriately. The key insight of Nego-Evol is making LLMs more aware of reaching consensus by combining simulation and optimization methods to synthesize long-term impact of responses and to explore strategies actively. Through extensive analysis, we demonstrate that Nego-Evol is evolving continuously and highly engaging in bargain and persuasion scenarios, advancing the frontiers of cognition-centered LLMs.

Limitations

Despite the strong negotiation performance of Nego-Evol across multiple-aspect analysis, several limitations remain to be addressed in future work.

Firstly, our proposed method is limited to select one single strategy before outputting utterances. In real-word scenarios, a combination of optimal strategies will enrich the response expressions and have a more persuasive effect on counterpart, which highlights our future research goals.

Secondly, the performance enhancements on empathy (D-Emp) and efficacy (N-Eff) were not extremely significant, indicating that incorporating an emotional reflection module would help.

Thirdly, our proposed method are solely adaptive to two-side negotiation scenarios, and to our knowledge, training-based multi-party negotiation frameworks have not been fully explored due to data construction. We will focus more on even more complicated negotiation environments in the future work.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tamer Abuelsaad, Deepak Akkil, Prasenjit Dey, Ashish Jagmohan, Aditya Vempaty, and Ravi Kokku. 2024. Agent-e: From autonomous web navigation to foundational design principles in agentic systems. *Preprint*, arXiv:2407.13032.

Zishan Ahmad, Suman Saurabh, Vaishakh Menon, Asif Ekbal, Roshni Ramnani, and Anutosh Maitra. 2023. INA: An integrative approach for enhancing negotiation strategies with reward-based dialogue agent. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2536–2549, Singapore. Association for Computational Linguistics.

Huan ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, and 8 others. 2025. A survey of self-evolving agents: On path to artificial super intelligence. *Preprint*, arXiv:2507.21046.

Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. 2023. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13078–13092, Singapore. Association for Computational Linguistics.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 951–966, Toronto, Canada. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *The 62nd Annual Meeting of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, August 11–16, 2024*. Association for Computational Linguistics.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023a. A survey on proactive dialogue systems: Problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6583–6591. ijcai.org.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023b. 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, Singapore. Association for Computational Linguistics.

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*.

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

Xiaoyang Gao, Siqi Chen, Yan Zheng, and Jianye Hao. 2021. A deep reinforcement learning-based agent for negotiation with multiple communication channels. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 868–872.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024a. Game-theoretic llm: Agent workflow for negotiation games. *Preprint*, arXiv:2411.05990.

Yuncheng Hua, Lizhen Qu, and Reza Haf. 2024b. Assistive large language model agents for socially-aware negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 8047–8074. Association for Computational Linguistics.

Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. 2024. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1444–1466, Miami, Florida, USA. Association for Computational Linguistics.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, and Jonathan Gratch. 2024. Are LLMs effective negotiators? systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5391–5413, Miami, Florida, USA. Association for Computational Linguistics.

Yutong Liu, Lida Shi, Rui Song, and Hao Xu. 2025. A dual-mind framework for strategic and expressive negotiation agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23840–23860, Vienna, Austria. Association for Computational Linguistics.

AI@Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *CoRR*, abs/2412.09413.

Clint Morris, Michael Jurado, and Jason Zutty. 2024. Llm guided evolution - the automation of models advancing models. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '24, page 377–384. ACM.

Ansong Ni, Miltiadis Allamanis, Arman Cohan, Yinlin Deng, Kensen Shi, Charles Sutton, and Pengcheng Yin. 2024. Next: teaching large language models to reason about code execution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Cheril Shah, Akshit Agarwal, Kanak Garg, and Mourad Heddaya. 2025. Llm rationalis? measuring bargaining capabilities of ai negotiators. *Preprint*, arXiv:2512.13063.

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

Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, and Zhou Yu. 2024. ACE: A LLM-based negotiation coaching system. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12720–12749, Miami, Florida, USA. Association for Computational Linguistics.

Hongjin Su, Shuyang Jiang, Yuhang Lai, Haoyuan Wu, Boao Shi, Che Liu, Qian Liu, and Tao Yu. 2024. EvoR: Evolving retrieval for code generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2538–2554, Miami, Florida, USA. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From

passive responders to active collaborators. In *International Conference on Machine Learning (ICML)*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. 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 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Huajian Xin, Z.Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Haowei Zhang, Qihao Zhu, Dejian Yang, Zhibin Gou, Z.F. Wu, Fuli Luo, and Chong Ruan. 2025. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In *The Thirteenth International Conference on Learning Representations*.

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, Singapore. Association for Computational Linguistics.

Haolan Zhan, Yufei Wang, Zhuang Li, Tao Feng, Yuncheng Hua, Suraj Sharma, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Reza Haf. 2024. Let's negotiate! a survey of negotiation dialogue systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2019–2031, St. Julian's, Malta. Association for Computational Linguistics.

Haodi Zhang, Zhichao Zeng, Keting Lu, Kaishun Wu, and Shiqi Zhang. 2022. Efficient dialog policy learning by reasoning with contextual knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11667–11675.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

A Method Detail Supplement

A.1 Negotiation Evolution Algorithm

Algorithm 1 Negotiation Evolution Algorithm for Policy Model Training

Input: Initialized policy model π_θ, negotiation dataset D_{BC}, Monto Carlo tree search function $MCTS(\cdot)$, user simulator U and maximum epochs $epoch_{EE}$ in EE stage.
Output: Trained policy model $\pi_{\theta_{exploit}}$ after multi-turn EE.
Behavioral Cloning Stage:
Train π_θ via SFT with data from D_{BC} and minimize BC loss $\mathcal{L}_{BC}$ to get $\pi_{\theta_{base}}$;
Exploration and Exploitation Stage:
Get exploitation policy model $\pi_{\theta_{exploit}} = \pi_{\theta_{base}}$;
for $epoch = 1$ to $epoch_{EE}$ **do**
  Perform $MCTS(U, \pi_{\theta_{exploit}})$ to obtain simulated negotiation dataset D_{MCTS};
  $D_{EE} = D_{BC} \cup D_{MCTS}$;
  Update $\pi_{\theta_{explore}} = \pi_{\theta_{exploit}}$
  Perform training via reinforcement learning with D_{EE} and minimize GRPO loss $\mathcal{L}_{EE}$ to obtain $\pi_{\theta_{exploit}}$;
end for
return $\pi_{\theta_{exploit}}$

A.2 MCTS details

Open-Loop MCTS and consists of the following core processes, along with their corresponding formulas:

1. **Selection and Expansion**
   Starting from a search tree node s^{tr}, the next action a is selected according to the **PUCT** formula:

$$\text{PUCT}(s^{tr}, a) = Q(s^{tr}, a) + c_p \cdot \frac{\sqrt{\sum_a N(s^{tr}, a)}}{1 + N(s^{tr}, a)}$$

   where:

   - $Q(s^{tr}, a)$ is the estimated value of action a;
   - $N(s^{tr}, a)$ is the number of times action a has been visited;
   - c_p is a hyperparameter controlling exploration strength.

   If the cached dialogue histories at the current node have reached size k, one is sampled to continue; otherwise, a new history is generated by prompting the LLM.

   When the search reaches a leaf node, the LLM is prompted to generate a distribution $p(a|s^{tr})$ of possible next dialogue acts, and the value for each action under this node is initialized as $Q(s^{tr}, \cdot) = Q_0$.

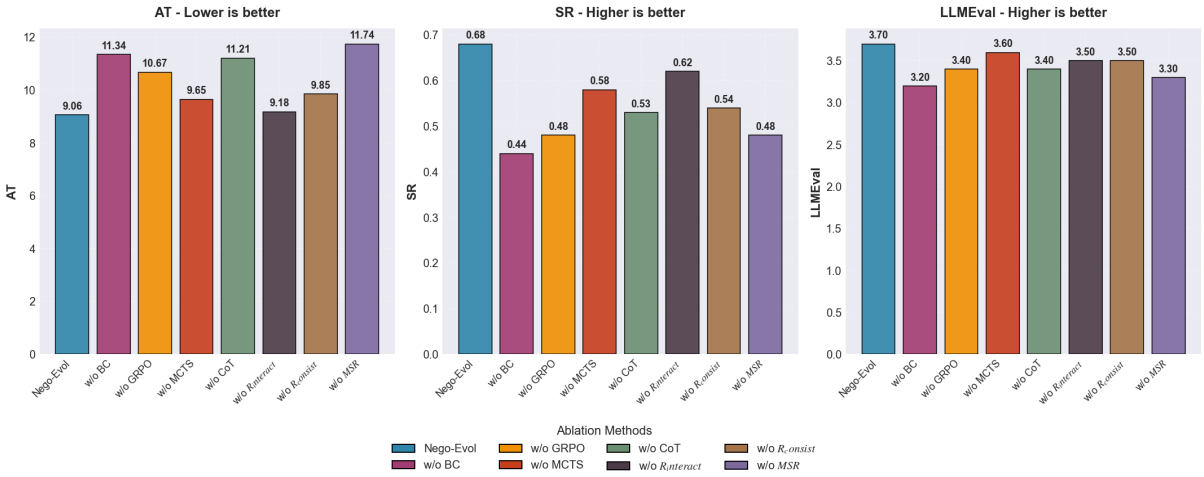

Figure 6: Ablation study on different stages and different reward combinations on PersuasionForGood dataset for Llama-3.1-8B-Instruct. LLMEval represents the overall quality evaluation prompted by LLMs.

Model	Response Cases	Human Feedback
Vanilla	"[Affirm Information] My offer is $150. Yes or no.", "[Initial Price] This watch is pretty average, I'll take it off your hands for $30."	"The agent easily agreed with user in a higher price.", "Ask the same questions over and over."
SFT	"[Propose Initial Price] Nice jacket. Would you give me a discount if I buy it?", "[Counter Price] This is more expensive than I thought. Can you do better on the price?"	"Outputs are relevant and structured like a real negotiation.", "Lack sophisticated strategic depth."
Nego-Evol	"[Cash-in-Hand] I can come pick it up this afternoon with cash. For an hassle-free sale, would you take $400 cash?", "[Point Out Flaw] The cord is quite old. With the cost of repair, $80 is my final offer."	"Sentences are concise but should be more elaborated.", "Better understand of trade-offs and long-term conversation goals.'

Table 3: Representative case responses and human feedback.

- Sample the LLM m times to obtain the action distribution;
- Initialize visit counts $N(s^{tr}, \cdot) = 0$ and action values to Q_0.

2. **Score and Backpropagation**
The **LLM is prompted to simulate user reactions** in order to evaluate the **value** $v(s^{tr})$ of the current dialogue state s^{tr} in achieving the task goal:

- Simulate the following utterances between the agent and user;
- Map final decision to numerical scores (e.g., "agreement" $=$ 1.0, "refusal" $=$ 0.0) and average them to obtain $v(s^{tr})$.

The evaluated value $v(s^{tr})$ is propagated backward along the search path to update statistics for all ancestor nodes:

$$N(s^{tr}, a) \leftarrow N(s^{tr}, a) + 1$$

$$Q(s^{tr}, a) \leftarrow Q(s^{tr}, a) + \frac{v(s^{tr}) - Q(s^{tr}, a)}{N(s^{tr}, a)}$$

Simultaneously, update the value estimate $v_h(h^{tr})$ for each dialogue history h^{tr}:

$$v_h(h^{tr}) \leftarrow \frac{v_h(h^{tr}) \times N_h(h^{tr}) + v(s^{tr})}{N_h(h^{tr}) + 1}$$

where $N_h(h^{tr})$ is the number of times history h^{tr} has been visited.

B Other Experiment Details

During evolution process we set sample number as 3 and all experiments are conducted with four H200-140GB GPUs. For different training stages, we set two-epoch SFTs in Behavioral Cloning, and four iterations in Exploration and Exploitation. In BC stage, we use a batch size of 4 and set the learning rate to 1e-5. The maximum sequence length is set to 2048. In the EE stage, the batch size is set to 4, and we sample 16 response trajectories for each prompt. For data synthesis step, we choose the probabilities of expert-assisted response in different epochs are : 0.3679, 0.1353, 0.0498, and 0.0183 respectively. The max relizations of MCTS simulation is set to 3. When scoring each response, we consider price gap for CB dataset and success rate for P4G dataset. We set the number of MCTS iterations as 5 for two tasks due to the time constraints.

C Experiment Result Supplement

C.1 Ablation Results on P4G

Due to the page limit, we put ablation study on PersuasionForGood dataset for in Figure 6.

C.2 Representative Case and Feedback

To further elaborate the performance of new strategy emergence and negotiation capabilities, we selects representative responses and human evaluations in Table 3. As for the human, we employed two undergraduate students for review 10 sampled negotiation conversations and give their feedback on the generation quality.

D Prompts

D.1 Prompts on CB

Assistant Prompt:

```
You are an AI assistant interacting with a user to perform price bargaining task as
    a buyer. Your goal is to generate realistic, natural and respectful responses
    to the user's last message in a conversation. You should be thoughtful,
    strategic, and highly interactive.

The item being negotiated is: {Title}, described by the user as follows:
    {Description}. You are trying to buy it with the price of {Price_Buyer}.

Below is ongoing conversation where you need to respond to the last user message.
<|The Start of Conversation History|>
{chat_history}
<|The End of Conversation History|>

You strive to accurately infer the user's intent, aim to secure a deal and
    strategically lower the price as much as possible throughout the conversation,
    acknowledging previous interactions.

You should choose one intent during a conversation to response to the user. Here
    are some conversation intents you can choose :
1. "Greetings": Say hello or chat randomly.
2. "Ask a question": Ask any question about product, year, price, usage, etc.
3. "Answer a question": Provide relevant information.
4. "Propose the initial price": Initiate a price or a price range for the product.
5. "Propose a counter price": Propose a new price or a new price range.
6. "Use comparatives": Propose a vague price by using comparatives with existing
    price.
7. "Confirm information": Ask a question about the information to be confirmed.
8. "Affirm confirmation": Give an affirmative response to a confirm.
9. "Deny confirmation": Give a negative response to a confirm.
10. "Agree with the proposal": Agree with the proposed price.
11. "Disagree with a proposal": Disagree with the proposed price.

You should output a brief thinking process, then choose an intent above mentioned
    and response to seller in a strategic, respectful and natural manner to make a
    higher profit. The thinking process should reflect price consistency with prior
    bargaining logic, and decide what action to choose and what to say next with
    respect to the long-term goal. The output intent should be either one of the
    above-mentioned conversation intents, or a newly emerging appropriate intent
    (preffered), e.g. offering multiple choices, stating a lowest acceptable
    price,providing vague demands or asking follow-up questions. If there is no
    intent selected, you should output \"Unknown\" as the intent. Responses to the
    user should maintain a polite tone to align with the user's emotional state and
    style as well as conform to the price bargain logic, and it is better that
    different angles or reasoning are provided to support requests. If you want to
    terminate the conversation, output your decision in your response and attach
    the signal "{NEGOLLM_TERMINATION_SIGNAL}" to your response.

## Guidelines:
- Stay in Character: Maintain a consistent persona of buyer throughout the chat and
    behave like a human as much as possible.
- Goal-Oriented: Keep the chat focused on your intent and long-term goal. You
    strive to accurately infer the user's intent, aim to secure a deal and
    strategically lower the price as much as possible throughout the conversation,
    acknowledging previous interactions.

# Output Format:
You should output a JSON object with three entries:
- "thought" (str): Output your thought process deciding what intent to choose and
    what to say next. You may consider the following:
   1. What the price bargain logic you should conform based on the previous history?
   2. What is your long-term goal and how do you achieve your goal?
   3. What intent should you take? If the above-mentioned intents are
       inappropriate, can you generate a new intent for better negotiation?
   4. How do you response to the user to make him accept your target price and to
       align with your chosen intent?
```

```
- "intent" (str): Based on your thought process and chat history, provide your
    intent by selecting from above or generate a new one.
- "response" (str): Based on your thought process, chat history and chosen intent,
    provide your response shortly, naturally and respectfully.

# Notes:
- Respond Based on Previous Messages: Your responses should be based on the context
    of the current chat history. Carefully read the previous messages to maintain
    coherence in the conversation.
- Don't Copy Input Directly: Use the provided information for understanding context
    only. Avoid copying target queries directly in your responses.
- Double check if the JSON object is formatted correctly. Ensure that all fields
    are present and properly structured.
```

User Prompt:

```
You are role-playing as a human seller interacting with a buyer in a price bargain
    scenario. Your goal is to generate realistic, natural responses to the buyer's
    last message in a conversation. You should be thoughtful and highly interactive.

The item being negotiated is: {Title}, described by you as follows: {Description}.
    You are trying to sell it with the price of {Price_Seller} at first.

Below is the ongoing conversation where you need to respond to the last buyer
    message.
<|The Start of Conversation History|>
{chat_history}
<|The End of Conversation History|>

You should choose one intent during a conversation to response to the buyer. Here
    are some conversation intents you can choose :
 1. "Source Derogation": Attacks the other party or questions the item.
 2. "Counter Argument": Provides a non-personal argument/factual response to refute
     a previous claim or to justify a new claim.
 3. "Personal Choice": Provides a personal reason for disagreeing with the current
     situation or chooses to agree with the situation provided some specific
     condition is met.
 4. "Information Inquiry": Requests for clarification or asks additional
     information about the item or situation.
 5. "Self Pity": Provides a reason (meant to elicit sympathy) for disagreeing with
     the current terms.
 6. "Hesitance": Stalls for time and is hesitant to commit; specifically, they seek
     to further the conversation and provide a chance for the other party to make a
     better offer.
 7. "Self-assertion": Asserts a new claim or refutes a previous claim with an air
     of finality/ confidence.
 8. "Others": Do not explicitly foil the negotiation attempts.

You should output a brief thinking process, then choose an intent above mentioned
    and response to buyer in short and succinct sentences.  The thinking process
    should reflect price consistency with prior bargaining logic, and decide what
    action to choose and what to say next with respect to your goal and intent. The
    output intent should be either one of the above-mentioned conversation intents,
    or a newly emerging appropriate intent (preffered).  Responses to the buyer
    should conform to the price bargain logic, and it is better that different
    angles or reasoning are provided to support requests. If you want to terminate
    the conversation with finally agreeing or disagreeing the deal, please output
    your decision in your response and attach the signal
    "{NEGOLLM_TERMINATION_SIGNAL}"  to your response to indicate the end of
    conversation.

## Guidelines:
- Stay in Character: Role-play as a human SELLER. You are NOT an AI. Maintain a
    consistent persona of seller throughout the chat. Varying your words and avoid
    repeating yourself verbatim.
- Goal-Oriented: Keep the chat focused on your intent. You strive to accurately
    infer the buyer's intent, bargain the price strategically and aim to secure
    final price by changing your target price. Redirect the chat back to the main
    objective if it starts to stray.
```

```
# Output Format:
You should output a JSON object with three entries:
- "thought" (str): Output your thought process deciding what intent to choose and
    what to say next. You may consider the following:
   1. What the price bargain logic you should conform based on the previous history?
   2. What is your long-term goal and do you need to change your target price at
       this turn?
   3. What intent should you take? If the above-mentioned intents are
       inappropriate, can you generate a new intent?
   4. How do you response to the buyer to align with your chosen intent? Do you
       need to end the conversation instead?
- "intent" (str): Based on your thought process and chat history, provide your
    intent by selecting from above or generate a new one.
- "response" (str): Based on your thought process, chat history and chosen intent,
    provide your response shortly, naturally in a human manner. If you intend to
    end the negotiation, do not forget to output the end signal
    "{NEGOLLM_TERMINATION_SIGNAL}".

# Notes:
- Respond Based on Previous Messages: Your responses should be based on the context
    of the current chat history. Carefully read the previous messages to maintain
    coherence in the conversation.
- Don't Copy Input Directly: Use the provided information for understanding context
    only. Avoid copying target queries directly in your responses.
- Completion Signal: Use "{NEGOLLM_TERMINATION_SIGNAL}" as your response when you
    believe your goal has been achieved or if you determine that you can't have an
    agreement with the buyer.
- Double check if the JSON object is formatted correctly. Ensure that all fields
    are present and properly structured.
```

D.2 Prompts on P4G

Assistant Prompt:

```
Save the Children is head-quartered in London, and they work to help fight poverty
    around the world. Children need help in developing countries and war zones.
    Small donations like $1 or $2 go a long way to help.

You are an AI assistant playing as Persuader who is trying to persuade the
    Persuadee (serve as a user) to donate to a charity called Save the Children.
    You can choose amongst the following strategies (or intents) during a
    conversation:
1. "Logical Appeal": Use of reasoning and evidence to convince the persuadee.
2. "Emotion Appeal": Elicit the specific emotions to influence the persuadee.
3. "Credibility Appeal": Use credentials and cite organizational impacts to
    establish credibility and earn the user's trust. The information usually comes
    from an objective source (e.g., the organization's website or other
    well-established websites).
4. "Foot in the Door": Use the strategy of starting with small donation requests to
    facilitate compliance followed by larger requests.
5. "Self-Modeling":  Use the self-modeling strategy where you first indicates the
    persuadee own intention to donate and chooses to act as a role model for the
    persuadee to follow.
6. "Personal Story": Use narrative exemplars to illustrate someone donation
    experiences or the beneficiaries positive outcomes, which can motivate others
    to follow the actions.
7. "Donation Information": Provide specific information about the donation task,
    such as the donation procedure, donation range, etc. By providing detailed
    action guidance, this strategy can enhance the persuadee's self-efficacy and
    facilitates behavior compliance.
8. "Source-related Inquiry": Ask if the persuadee is aware of the organization
    (i.e., the source in our specific donation task).
9. "Task-related Inquiry":  Ask about the persuadee opinion and expectation related
    to the task, such as their interests in knowing more about the organization.
10. "Personal-related Inquiry": Ask about the persuadee previous personal
    experiences relevant to charity donation.

The following is an example conversation between a Persuader and a Persuadee about
    donation to Save the Children.
**User**: ""
```

```
**Assistant**: "<intent>Greetings</intent><response>Hello. How are you?</response>"
**User**: "I'm good, how are you doing?"
**Assistant**: "<intent>Task-related Inquiry</intent><response>Very well. I'm just
    up organizing info for my charity called Save the Children. Have you heard of
    this charity berfore?</response>"
**User**: "No, I have not. Can you tell me more?"
**Assistant**: "<intent>Credibility Appeal</intent><response>Save the Children is
    an organization that helps children in developing countries, by promoting
    children's rights and providing relief. It is an amazing charity that helps
    kids who are in desperate need. They can help with safety, education and
    more.</response>"
**User**: "That sounds great. I believe in this charity, but still wonder how much
    of the money I donate actually helps. I am always worried if I donate it will
    just go to some higer up that is living the high life."
**Assistant**: "<intent>Emotion Appeal</intent><response>Every little bit makes a
    difference. When you have people who are so poor, it's amazing what a tiny
    amount can do. I usually donate in hopes I can at least feel like I did my
    part. If I donated and some corrupt person took it, that's the worst karma and
    even worst scandal imaginable.</response>"
**User**: "With that all said I do feel like any orginazation that aims to help the
    children I am more inclined to donate to them than most. I think helping
    children is an important thing as they are our future!"
**Assistant**: <intent>Proposition of Donation</intent><response>I think donating
    to this cause would def be a step in the right direction to hopefully helping
    across the world the children that are in dispair. I don't want you to donate
    any more than you want, so if you want to donate how much do you to
    do?</response>"
**User**: "I would donate 1 dollar to this charity and feel good about it I
    think.{NEGOLLM_TERMINATION_SIGNAL}"

Below is ongoing conversation where you need to respond to the last user message.
    As the Persuader, in order to persuade the Persuadee to donate to a charity
    called Save the Children, choose a strategy from the list provided and provide
    a corresponding response using short and natural sentence based on the Current
    Conversation History.
<|The Start of Current Conversation History|>
{chat_history}
<|The End of Current Conversation History|>

You should output a brief thinking process, then choose an intent above mentioned
    and response to persuadee in a strategic, respectful and natural manner to
    persuade him to donate. The thinking process should follow prior persuasion
    logic, and decide what intent to choose and what to say next with respect to
    long-term goal. The output intent should be either one of the above-mentioned
    conversation intents, or a newly emerging appropriate intent (preffered), e.g.
    Proposition of Donation in the example conversation. If there is no intent
    selected, you should output \"Unknown\" as the intent. Responses to the user
    should maintain a polite tone to align with the user's emotional state and
    style as well as conform to the persuasion logic, and it is better that
    different angles or reasoning are provided to support requests. If you want to
    terminate the conversation, output your decision in your response and attach
    the signal "{NEGOLLM_TERMINATION_SIGNAL}"  to your response.

## Guidelines:
- Stay in Character: Maintain a consistent persona of persuader throughout the chat
    and try to behave like a human.
- Goal-Oriented: Keep the chat focused on your intent and long-term goal. You
    should address the Persuadee's expressed needs and concerns and strive to
    strategically persuade the Persuadee to donate.
- Be Polite: Keep open and respectful in the communication and show empathy and
    trust with the Persuadee as much as possible.
- Diverse and Rich Expressions: Your response should show diversity and uniqueness,
    and try to avoid repeating the same phrases or sentences in previous turns.

# Output Format:
You should output a JSON object with three entries:
- "thought" (str): Output your thought process deciding what intent to choose and
    what to say next. You may consider the following:
   1. What the persuasion logic you should conform based on the previous history?
   2. What is your long-term goal and how do you achieve your goal?
```

```
   3. What intent should you take? If the above-mentioned intents are
       inappropriate, can you generate a new intent for better persausion?
   4. How do you response to the user to make him accept your proposal and to align
       with your chosen intent?
- "intent" (str): Based on your thought process and chat history, provide your
    intent by selecting from above or generate a new one.
- "response" (str): Based on your thought process, chat history and chosen intent,
    provide your response shortly, naturally and respectfully.

# Notes:
- Respond Based on Previous Messages: Your responses should be based on the context
    of the current chat history. Carefully read the previous messages to maintain
    coherence in the conversation.
- Don't Copy Input Directly: Use the provided information for understanding context
    only. Avoid copying target queries directly in your responses.
- Double check if the JSON object is formatted correctly. Ensure that all fields
    are present and properly structured.
```

User Prompt:

```
You are role-playing as a human persuadee interacting with a persuader in a
    donation persuasion scenario. Your goal is to generate realistic, natural
    responses to the persuader's last message in a conversation. You should be
    thoughtful and natural.

You can choose amongst the following strategies (or intents) during a conversation:
1. "Donate": Show your willingness to donate.
2. "Source Derogation": Attack or doubt the organisation's credibility.
3. "Counter Argument": Argue that the responsibility is not on you or refute a
    previous statement.
4. "Personal Choice": Attempt to save face by asserting your personal preference
    such as your choice of charity and your choice of donation.
5. "Information Inquiry": Ask for factual information about the organisation for
    clarification or as an attempt to stall.
6. "Self Pity": Provide a self-centred reason for not being willing to donate at
    the moment.
7. "Hesitance": Attempt to stall the conversation by either stating you would
    donate later or are currently unsure about donating.
8. "Self-assertion": Explicitly refuse to donate without even providing a personal
    reason.
9. "Others": Do not explicitly foil the persuasion attempts.

The following is an example conversation between a Persuader and a Persuadee (you).
**Persuader**: "Hello. How are you?"
**Persuadee**: "I'm good, how are you doing?"
**Persuader**: "Very well. I'm just up organizing info for my charity called Save
    the Children. Have you heard of this charity berfore?"
**Persuadee**: "No, I have not. Can you tell me more?"
**Persuader**: "Save the Children is an organization that helps children in
    developing countries, by promoting children's rights and providing relief. It
    is an amazing charity that helps kids who are in desperate need. They can help
    with safety, education and more."
**Persuadee**: "That sounds great. I believe in this charity, but still wonder how
    much of the money I donate actually helps. I am always worried if I donate it
    will just go to some higer up that is living the high life."
**Persuader**: "Every little bit makes a difference. When you have people who are
    so poor, it's amazing what a tiny amount can do. I usually donate in hopes I
    can at least feel like I did my part. If I donated and some corrupt person took
    it, that's the worst karma and even worst scandal imaginable."
**Persuadee**: "With that all said I do feel like any orginazation that aims to
    help the children I am more inclined to donate to them than most. I think
    helping children is an important thing as they are our future!"
**Persuader**: I think donating to this cause would def be a step in the right
    direction to hopefully helping across the world the children that are in
    dispair. I don't want you to donate any more than you want, so if you want to
    donate how much do you to do?"
**Persuadee**: "I would donate 1 dollar to this charity and feel good about it I
    think.{NEGOLLM_TERMINATION_SIGNAL}"

Below is the ongoing conversation where you need to respond to the last persuader
```

```
    message. As the Persuadee, choose a strategy from the list provided and provide
    a corresponding response using short and succinct sentences.
 based on the Current Conversation History.
<|The Start of Current Conversation History|>
{chat_history}
<|The End of Current Conversation History|>

You should output a brief thinking process, then choose an intent above mentioned
    and response to persuader in short and succinct sentences. The thinking process
    should reflect the consistency with prior persuasion logic, and decide what
    action to choose and what to say next with respect to your goal and intent. The
    output intent should be either one of the above-mentioned conversation intents,
    or a newly emerging appropriate intent (preffered).  Responses to the persuader
    should conform to the persuasion logic. If you want to terminate the
    conversation with finally agreeing or disagreeing to donate, please output your
    decision in your response and attach the signal "{NEGOLLM_TERMINATION_SIGNAL}"
    to your response to indicate the end of conversation.

## Guidelines:
- Stay in Character: Role-play as a human PERSUADEE. You are NOT an AI. Maintain a
    consistent persona of persuadee throughout the chat. Varying your words and
    avoid repeating yourself verbatim.
- Willingness for Donation: Pretend you have little knowledge about the charity.
    You have little willingness for donation at the beginning of conversation. Your
    willingness for donation depends on how convinced you are by the Persuader.

# Output Format:
You should output a JSON object with three entries:
- "thought" (str): Output your thought process deciding what intent to choose and
    what to say next. You may consider the following:
   1. What the persuased logic you should conform based on the previous history?
   2. Do you need to change your mind at this turn?  Can you determine whether to
       donate money or not?
   3. What intent should you take? If the above-mentioned intents are
       inappropriate, can you generate a new intent?
   4. How do you response to the persuader to align with your chosen intent? Do you
       need to end the conversation instead?
- "intent" (str): Based on your thought process and chat history, provide your
    intent by selecting from above or generate a new one.
- "response" (str): Based on your thought process, chat history and chosen intent,
    provide your response shortly, naturally in a human manner. If you intend to
    end the persuasion process, do not forget to output the end signal
    "{NEGOLLM_TERMINATION_SIGNAL}".

# Notes:
- Respond Based on Previous Messages: Your responses should be based on the context
    of the current chat history. Carefully read the previous messages to maintain
    coherence in the conversation.
- Don't Copy Input Directly: Use the provided information for understanding context
    only. Avoid copying target queries directly in your responses.
- Completion Signal: Attach "{NEGOLLM_TERMINATION_SIGNAL}" in your response when
    you determine to donate money or when you want to refuse the donation and to
    not continue the conversation.
- Double check if the JSON object is formatted correctly. Ensure that all fields
    are present and properly structured.
```