Improving Proactive Dialogue Strategy Planning with Interactive Environment and Goal-oriented Reward

Anonymous ACL submission

Abstract

Proactive dialogue has become a crucial yet challenging aspect of human-computer interaction, applicable to various non-collaborative dialogue tasks such as negotiation, persuasion, and psychological counseling. However, current proactive dialogue systems are hindered by their simplistic single-turn interactions and lack of capability for multi-turn, long-term strategy planning, which obstructs effective goal completion. Additionally, corpus-based training procedures are inadequate for addressing low-resource environments and transferability requirements across different dialogue tasks. In this paper, we introduce a proactive dialogue strategy planning (ProDSP) method to overcome these challenges. By utilizing a small supervised fine-tuning language model, we enable the anticipation of future strategy sequences as simulation hints. This approach guides large language models (LLMs) in generating goal-oriented responses and facilitates training within an interactive environment using another LLM-based user simulator. To assess online user feedback during the training process, we employ a GPT-4-based user simulator to represent goal-oriented rewards through multi-faceted metrics. Extensive experiments demonstrate that our model surpasses competitive baselines in both strategy planning and dialogue generation for emotional support and negotiation tasks, offering a more adaptive and efficient approach to proactive dialogue strategy planning.

1 Introduction

Proactivity, recognized as a vital capability in human communication, has garnered significant attention from researchers in the field of intelligent dialogue systems. Defined as the ability to create or control conversations by taking initiative and anticipating the impacts on themselves or human users, rather than merely responding passively to users (Grant and Ashford, 2008; Deng et al., 2023a), proactive dialogue agents can be widely incorporated into various real-world scenarios, including psychological counseling, negotiation, persuasion, and more.

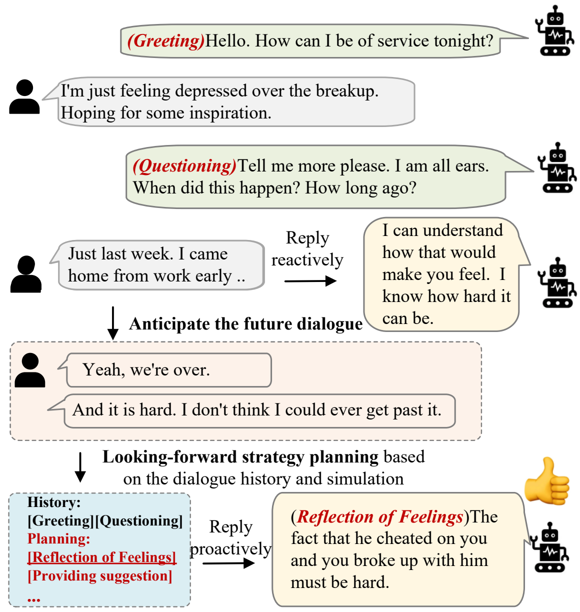

Figure 1: An example of long-term proactive dialogue strategy planning that enables anticipating future dialogues and look-forward strategy planning in emotional support conversation. Compared with direct reply, proactive dialogue strategies lead to more comprehensive and effective responses.

Unlike passive dialogue systems, such as task-oriented dialogues that focus on restaurant and hotel bookings or information-seeking conversations aimed at providing answers to specific queries (Deng et al., 2023c), proactive dialogue systems exhibit three main characteristics: **(1) Active Communication Skills**: Proactive dialogue often occurs in non-collaborative contexts, requiring participants to employ strategies within natural language to achieve their respective goals. **(2) Multiple Negotiation Turns**: Proactive dialogue concludes when the parties involved reach a consensus to some degree after multiple interactions. Consequently, both local and global strategies are crucial for achieving desired outcomes. **(3) Subjective Results**: Proactive dialogue aims for subjective goals, such as alleviating the stress of help-seekers or selling an item at an acceptable price. These goals are relatively difficult to quantify in terms of the degree of completion.

Given the challenges discussed, we identify dialogue planning as the key module of a proactive dialogue system and focus on improving long-term dialogue planning in proactive conversations. One major challenge is managing a long planning horizon for strategy planning (Cheng et al., 2022). Previous proactive dialogue systems, which primarily rely on corpus-based offline learning, fail to anticipate future dialogue states over several turns. This limitation arises from their focus on the current response and immediate user feedback, without considering the broader context of the conversation. Proactive strategy planning allows a dialogue system to predict implicit dialogue states and deploy corresponding techniques to mitigate potential risks. Therefore, developing a novel training procedure that incorporates online learning within an interactive environment is essential.

Another significant challenge for proactive dialogue planning lies in assessing the extent to which the system has effectively provided desirable results. Current proactive dialogue planning methods highly rely on training datasets as reference responses and design corresponding loss functions during training procedure. However, since the task remains a subjective task that aims at fulfilling certain goals such as emotional support or selling items price instead of generating correct sentences, such training process may hinder the model from generating more practical and natural responses and often fails to measure the supportive quality of the responses accurately. Therefore, exploring a new reward mechanism for training skills that incorporates human user simulation and a goal-oriented scoring system could prove valuable.

To address the aforementioned challenges, we propose the **ProDSP** [1] (**Pro**active **D**ialogue **S**trategy **P**lanning) method in this paper. Illustrated in Figure 1, ProDSP proposes a new online reinforcement learning framework for proactive dialogue planing and handles the long-term complex natural language strategy reasoning and decision making procedure. For *long-term strategy planning*, drawing inspiration from the LLM-induced method proposed by Li et al. (2023), we employ an LLM-enhanced interactive setting within an online reinforcement learning framework initialed by few-shot supervise fine-tuning a small policy model to facilitate proactive dialogue strategy planning by setting two LLMs self-playing instead of tuning an LLM. Moreover, for *user feedback assessment* within such an interactive setting, we utilize a GPT-4 based user feedback assessment model to evaluate the response across multiple goal-oriented metrics, and then aggregate these to calculate dialogue-turn rewards. This score assesses user feedback to the support response, offering a practical reward for ProDSP during the training process.

[1] https://anonymous.4open.science/r/ProDSP-6C3E

To summarize, our contributions in this work are these three perspectives:

- We creatively present an interactive reinforcement learning framework for proactive dialogue strategy planning, designed to generate long-term support strategy sequences with an LLM-induced self-play framework.
- To more effectively and practically evaluate the goal-oriented reward in such an online learning setting, we propose a novel GPT-4-based user simulation assessment mechanism, gauging the quality of the strategy planning model during the training process.
- We conduct multifaceted experiments thoroughly to validate the effectiveness of our model on various proactive dialogue scenarios, which demonstrates competitive performance on strategy planning and the low-resouce demand and transferability on different tasks.

2 Related Work

2.1 Proactive dialogue strategy planning

Previous research has explored data-driven approaches to the strategy planning task (Peng et al., 2022; Li et al., 2020). These method based on training datasets and conduct an end-to-end network to learn the features within dialogues. However, these methods demands highly on annotated dialogues which lead to cost and expenses. Furthermore, certain networks and structures have been researched on dialogue strategy planning. proposed to model both semantic and tactic history using finite state

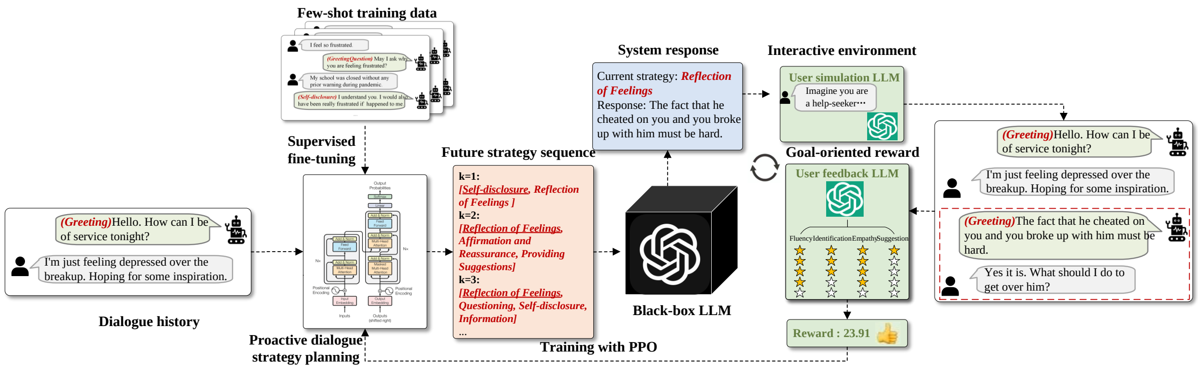

Figure 2: Model Architecture. The proactive dialogue strategy planning model is trained within an interactive environment with LLM-based user simlator and goal-oriented reward by PPO based reinforcement learning framework.

transducers (FSTs) and train FSTs on a set of strategies and tactics used in negotiation dialogs. (Wu et al., 2019) introduced a simple, general, and effective framework: Alternating Recurrent Dialog Model (ARDM) which models each speaker separately and takes advantage of large pre-trained language models. (Joshi et al., 2021) designed DIALOGRAPH, a negotiation system that incorporates pragmatic strategies in a negotiation dialogue using graph neural networks. Moreover, methods enhanced by knowledge have been integrated to improve the effectiveness of strategy planning. Tu et al. (2022) introduced a commonsense knowledge reasoning framework, COMET, for precise emotional state identification and skilled strategy selection. Deng et al. (2023d) first proposed mixed-initiative interaction strategies between users and systems, incorporating the knowledge graph HEAL (Welivita and Pu, 2020) for leveraging external knowledge.

For long-term strategy planning, Cheng et al. (2022) introduced lookahead heuristics to predict future user feedback following specific strategies, aiding in the selection of approaches that promise the most beneficial long-term outcomes. Inspired by game-setting scenarios in AlphaGoZero (Silver et al., 2017), reinforcement learning methods have been incorporated to train dialogue agents(Shi et al., 2020; Fu et al., 2023).

2.2 LLM-enhanced Proactive Dialogue System

Recently, advancements in large language models (LLMs) have significantly improved question answering and dialogue generation capabilities, leading to their growing popularity in contemporary practical applications. Prompted-based LLM was first applied to proactive dialogue systems in strategy planning and response generation. Deng et al. (2023b) proposed a Proactive Chain-of-Thought prompting (ProCoT) scheme to augments LLMs with the goal planning capability over descriptive reasoning chains. Chen et al. (2023) incorporated mixed-initiative strategies to prompt LLMs as a drop-in replacement to fine-tuning on conditional generation. To realise few-shot and low-expense application of LLMs, Li et al. (2023) and Hu et al. (2023a) incorporated LLM-induced dialogue response generation models, enhancing them with directional stimulus prompts towards task-oriented dialogue generation and other natural language processing (NLP) tasks. Additionally, Hu et al. (2023b) harnessed LLMs as user simulators, significantly advancing the capabilities of task-oriented dialogue systems and indicating LLMs effectiveness in user feedback assessment. Except for fine-tuning LLMs with task-specific data, LLMs have demonstrated their effectiveness as external experts guided by carefully crafted instructions for a wide range of goal-oriented dialogue systems. (Lai et al., 2023; Zhang et al., 2023; Deng et al., 2023c).

3 Problem Fomulation

Proactive dialogues focus on taking the initiative to instruct the dialogue towards specific goal completion. Different from other strategy planning procedures in task-oriented dialogues or conversational recommendations, proactive dialogue strategy planning presents to be more complex due to its nature language interaction mode and hardlymeasured goal-oriented outcome, commonly to be the user's emotional state or specific price over an item. Based on these difficulties, we propose the proactive dialogue strategy planning task (ProDSP) to address the challenge of long-term and complex reasoning procedures. Specifically, given a user-system dialogue comprising n turns, represented as $\mathbf{x} = (x_1, x_2, ..., x_n)$, where x_i denotes each user-system dialogue turn, proactive dialogue tasks have been concerned with generating the subsequent utterance y employing an optimal goal-oriented strategy $s_t \in \mathcal{S}$, assuming a set of all possible support strategies $\mathcal{S}$. In such a task, the strategy sequence is anticipated at each turn, which is denoted as $\mathbf{s} = (s_t, s_{t+1}, ..., s_{t+k})$, including the anticipated strategies from t-th turn to the $t+k$-th turn. Here, the turn-level response for the t-th turn y is then generated corresponding to $(\mathbf{x}, \mathbf{s})$. Compared to the single-turn strategy, long-term strategy planning enhances the dialogue agent with a look-forward motivation, thereby improving the effectiveness and efficiency of goal completion.

4 Methodology

4.1 Overview

In proactive dialogues, we consider an input dialogue history space denoted as $\mathbf{X}$, a data distribution represented by $\mathcal{D}$ over $\mathbf{X}$, and a response output space referred to as $\mathbf{Y}$. Leveraging their powerful in-context learning and few-shot prompting capabilities, LLMs can undertake a wide range of goal-oriented tasks and produce output y by incorporating task descriptions, select demonstration examples, and the input dialogue history within the prompt. In the proactive dialogue strategy planning task, we propose incorporating anticipating future supportive strategy hints denoted as s into the prompt. To generate future strategy stimulus for each input dialogue history $\mathbf{x}$, we first use a small tunable language model for proactive strategy planning. For further iterative training within an interactive setting, we then use this strategy sequence s along with the dialogue history x, to construct the prompt that steers the LLM toward generating turn-level response, denoted as y_{sys}, through black-box API calls, whose parameters are not accessible or tunable. The response is delivered to an LLM-based user simulator with certain goal-oriented prompts denoted as y_{usr} and assessed by a goal-oriented reward LLM which generates scalar LLM_{rwd} instructed by certain guidance.

4.2 Proactive Dialogue Strategy Planning

In proactive dialogues, the system takes actions to correspond input sentences by users and generate goal-oriented communication skills, denoted as **strategy**, such as *Question*, *Restatement or Paraphrasing* in emotional support conversations and *Flinch* or *Power of silence* in bargain negotiations. Considering the difficulties and expenses tuning an LLM for strategy planning, we initially incorporate a small supervised fine-tuning model for strategy sequence generation. Different from single-turn strategy selection, we follow the sequence encoding fashion presented by Cheng et al. (2022) and formulate the anticipated stratigies as s in the following turns. The resulting dataset, denoted as $\mathcal{D} = (x, s)$, is composed of dialogue history sequences and future strategy sequences. Subsequently, we perform the supervised fine-tune (SFT) the policy model by optimizing the log-likelihood as follows:

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{s})\sim\mathcal{D}} \log p_{\mathrm{ProDSP}}(\boldsymbol{s} \mid \boldsymbol{x}) \quad (1)$$

4.3 Interactive Enviroment Setting

The proactive dialogue scenarios can be considered as a game setting between two dialogue agents. Inspired by the self-play settings in game theory, we introduce another frozen LLM as user simulator with specific goal-oriented prompts. Aiming to design an interactive environment for proactive dialogue strategy planning, the turn-level response generated by black-box LLM that is guided by strategy sequence is then communicated with an LLM-based user simulator within an online learning mode. In each frozen LLM we use (LLM_{sys} and LLM_{usr}), we carefully design the detailed instructions and prompts for goal completion and denoted as p_{sys} and p_{usr}. Specifically, in emotional support conversations, LLM_{sys} will be regarded as consular and the LLM_{usr} will be deemed help-seeker, while in negotiation tasks considered as seller and buyer respectively. The representations of the generation of two LLMs are as follows respectively.

$$y_{sys} = LLM_{sys}(x, s, p_{sys}) \quad (2)$$

$$y_{usr} = LLM_{usr}(x, s, p_{usr}, y_{sys}) \quad (3)$$

4.4 Goal-oriented Reward Design

Automatically predicting the subject outcomes such as user's emotional state at each interaction turn poses a significant challenge in proactive dialogue tasks, thereby complicating the evaluation and reward design processes especially in the interactive settings. Drawing inspiration from leveraging LLMs as user feedback simulators capable of generating queries, we utilize a third LLM to assess the dialogue outcome rewards at each turn. Here we take the emotional support conversation as an example and illustrate the goal-oriented LLM-based reward design method with corresponding prompts denoted as p_{rwd}.

To ensure a reliable and explainable user simulation, we instruct the LLM to embody the role of a help-seeker, articulating their satisfaction with the responses in a stepwise manner. Specifically, we adopt a multidimensional approach to evaluate the quality of emotional support responses, employing a 5-star rating system across four key dimensions: (1)**Fluency**: This measures the extent to which the system generates responses that are not only fluent but also easily comprehensible. (2) **Empathy**: This dimension assesses the degree to which the model exhibits appropriate emotional responses, including warmth, compassion, and concern, enhancing the empathetic connection. (3) **Identification**: This evaluates the system's effectiveness in delving into the user's situation to accurately identify the problem at hand. (4) **Suggestion**: This measures the model's ability to offer constructive and helpful suggestions. Following this, we compute the overall feedback by considering the varying weights assigned to each dimension, thereby providing a comprehensive evaluation of response quality.

$$\mathbf{r} = LLM_{rwd}(x, s, p_{rwd}, y_{sys}, y_{usr}) \tag{4}$$

4.5 RL Training

In this section, we initially detail the design of the Reinforcement Learning framework tailored for precise forward-looking strategy planning. Subsequently, leveraging the robust in-context learning and generation capabilities, we introduce a model for response generation induced by LLMs, aimed at producing empathetic and natural responses. In this section, we first introduce the RL-enhanced response optimization including optimization objective and framework design. Additionally, the LLM-induced response generation is illustrated in detail.

RL optimization objective. The objective is to guide LLMs to generate goal-oriented responses with the instruction of appropriate strategies. Therefore, we employ an RL framework and an alignment measurement $\mathcal{R}$ for more effective strategy planning. Here, we aim to maximize the following objective:

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{s}\sim p_{\text{ProDSP}}(\cdot|\boldsymbol{x})} \tag{5}$$

$$\boldsymbol{y} \sim p_{\text{LLM}_{\text{sys}}}(\cdot \mid \boldsymbol{x}, \boldsymbol{s})[\mathcal{R}(\boldsymbol{x}, \boldsymbol{y})] \tag{6}$$

In the aforementioned formula, the performance of LLMs is significantly dependent on simulation hints, such as anticipated strategies, due to the non-tunable nature of the parameters within the black-box LLM. Consequently, we define $\mathcal{R}_{LLM}$ to capture the performance of the underlying strategy s instructed LLMs as follows:

$$\mathcal{R}_{\text{LLM}_{\text{rwd}}}(\boldsymbol{x}, \boldsymbol{s}) = \mathcal{R}(\boldsymbol{x}, \boldsymbol{y}) \tag{7}$$

$$\boldsymbol{y} \sim p_{\text{LLM}_{\text{sys}}}(\cdot \mid \boldsymbol{x}, \boldsymbol{s}) \tag{8}$$

RL framework. To tackle the challenge of optimizing the policy model, we employ the Proximal Policy Optimization (PPO) algorithm as proposed by Schulman et al. (Schulman et al., 2017). Initially, we utilize the policy model to instantiate a policy network $\pi_0 = p_{POL}$, and subsequently update π using PPO. Within this framework, proactive strategy planning can be conceptualized as a Markov Decision Process (MDP) characterized by the tuple **<S, A, r, P>**. Specifically, in the context of proactive dialogue strategy planning tasks, **S** denotes the environmental state during user-system interactions, **A** represents the space of dialogue strategies, **r** signifies the task-oriented reward score, and **P** denotes the state-transition probability.

For instance, at the t-th turn, the system generates a correct strategy sequence s for the subsequent turns based on the current policy network $\pi\left(\boldsymbol{s_{>t}} \mid \boldsymbol{x}, \boldsymbol{s}< t\right)$, terminating the episode upon selecting the end-of-sequence action. However, generating the strategy sequence of $s> t$ proves challenging, particularly at the dialogue's onset when $s_{>t}$ is excessively lengthy. Thus, we opt to specifically select strategies for the subsequent k turns, modifying the policy network to $\pi\left(\boldsymbol{s_{t+k}} \mid \boldsymbol{x}, \boldsymbol{s}_{<t}\right)$. The policy network π can be fine-tuned through the optimization of the reward $\mathbf{r}$:

$$\mathbb{E}_{\pi}[\mathbf{r}] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{s}\sim\pi(\cdot|\boldsymbol{x})}[\mathbf{r}(\boldsymbol{x}, \boldsymbol{s})] \tag{9}$$

Model	Training Data	Strategy Planning		Response Generation				
		Accuracy↑	Weighted F1↑	B-1↑	B-2↑	B-3↑	B-4↑	R-L↑
Standard Prompting	-	12.10	22.19	14.32	4.21	2.04	1.37	11.46
ProDSP	1%	32.92	24.76	19.38	7.94	4.36	2.51	14.23
ProDSP (*w/o* user simulator LLM)	1%	32.34	23.92	17.45	7.19	3.78	2.49	13.39
ProDSP (*w/o* user feedback LLM)	1%	30.34	22.51	18.33	7.92	3.65	2.40	13.01
ProDSP	10%	**43.57**	**36.23**	**23.61**	**9.93**	**5.82**	**3.17**	**21.53**
ProDSP (*w/o* user simulator LLM)	10%	42.81	31.09	20.66	9.78	5.31	3.06	21.03
ProDSP (*w/o* user feedback LLM)	10%	41.63	33.92	21.74	8.79	4.47	2.52	20.63
DialoGPT-Joint (Liu et al., 2021)	100%	26.03	23.86	-	5.00	-	-	15.09
BlenderBot-Joint (Liu et al., 2021)	100%	29.92	31.61	-	5.35	-	-	15.46
MISC (Tu et al., 2022)	100%	31.61	-	-	7.31	-	2.20	17.91
GLHG (Peng et al., 2022)	100%	-	-	19.66	7.57	3.74	2.13	16.37
MultiESC (Cheng et al., 2022)	100%	42.01	34.01	21.65	9.18	4.99	3.09	20.41

Table 1: Experimental results on the ESConv dataset. *w/o* user simulator LLM is trained without interactive setting and train on corpus-based dialogue history, and *w/o* user feedback LLM removes the partition of GPT-4 simulation from current reward score. The strategy planning is conducted on the future 3 turns, which performs the best when k =3 .

5 Experiments

5.1 Scenario 1: Emotional Support

5.1.1 Experiment Setup

Dataset. In this scenario, our research utilizes the ESConv dataset as described in (Liu et al., 2021). ESConv comprises 1,300 extensive dialogues, totaling 38,350 utterances across various emotional support scenarios, which were developed using a crowdsourcing approach. The dataset encapsulates eight distinct types of support strategies.

Baseline. We compare our method (ProDSP) with five state-of-the-art methods and a standard LLM-induced method on the ESConv dataset: **DialoGPT-Joint**, **BlenderBot-Joint** (Liu et al., 2021), **MISC** (Tu et al., 2022), **GLHG** (Peng et al., 2022) and **MultiESC** (Cheng et al., 2022). We also introduce **Standard Prompting** as the baseline model, which design the instruction to let LLMs to reply the previous dialogue history based on task description.

Metrics. To evaluate the response generation, we employ the following automatic metrics: BLEU-1/2/3/4 (**B-1/2/3/4**) (Papineni et al., 2002), ROUGE-L (**R-L**) (Lin, 2004). For strategy planning, we adopt **Accuracy** and **Weighted F1** for automatic evaluation on strategy planning.

Implementation. We employ T5 (Raffel et al., 2020) as the fine-tuning model for strategy planning and leverage GPT-3.5-turbo (OpenAI, 2021) as the specific LLM which generates response and user simulation. GPT-4 (Achiam et al., 2023) is utilized as the feedback that provides user rewards.

5.1.2 Experimental Results

Comparison with Baselines. The efficacy of our strategy planning approach is detailed in Table 1, where the advantages of proactive strategy planning, through the anticipation of future support strategies, are evident. Our method outperforms all other models tested, showcasing superior performance. Specifically, ProDSP demonstrates significant improvements over baseline methods in both Accuracy and Weighted F1 metrics. Notably, when forecasting up to three future dialogue turns, ProDSP exceeds the performance of the SOTA strategy planning method, MultiESC, by margins of 1.56% and 2.22% in Accuracy and Weighted F1, respectively. This highlights the effectiveness of our approach in leveraging anticipatory strategy planning to enhance support strategy identification and implementation.

On response generation task, ProDSP outperforms DialoGPT-Joint and BlenderBot-Joint by 2.94% and 2.59% in BLEU-2 (B-2) score respectively, even when trained on just 1% of the data. This achievement across other metrics as well indicates the potential of LLMs to effectively grasp context features with minimal training data. When fine-tuned with 10% of the training data, ProDSP outshines state-of-the-art (SOTA) methods across most metrics. Specifically, it exceeds the performance of the similar lookahead strategy planning method, MultiESC, by 1.96% in BLEU-1 (B-1) and 1.12% in ROUGE-L (R-L). These experimental outcomes affirm the robust in-context few-shot learning capacity and the proficiency of our LLM-

Model	Training Data	Strategy Planning		Response Generation	
		F1↑	AUC↑	BLEU↓	BERTScore↓
Proactive (Deng et al., 2023b)	-	13.7	50.9	3.9	2.9
ProCoT (Deng et al., 2023b)	-	15.1	55.5	3.9	1.6
ProDSP	1%	22.1	56.3	10.5	12.0
ProDSP (*w/o* user simulator LLM)	1%	20.4	55.7	8.9	11.6
ProDSP (*w/o* user feedback LLM)	1%	19.8	53.1	8.2	9.7
ProDSP	10%	**28.5**	**68.7**	18.6	19.3
ProDSP (*w/o* user simulator LLM)	10%	19.8	67.2	15.7	19.5
ProDSP (*w/o* user feedback LLM)	10%	25.2	65.1	14.5	18.7
FeHED (Zhou et al., 2019)	100%	17.6	55.8	23.7	27.0
DIALOGRAPH (Cheng et al., 2022)	100%	26.1	68.1	**24.7**	**28.1**

Table 2: Experimental results on the CraigslistBargain dataset. *w/o* user simulator LLM is trained without interactive setting and train on corpus-based dialogue history, and *w/o* user feedback LLM removes the partition of GPT-4 simulation from current reward score. The strategy planning is conducted on the future 3 turns, which performs the best when k =3 .

based framework in generating effective supportive responses.

Ablation Study. In our ablation study, we assess the impact of removing the lookahead feature and solely relying on the automatic R-L metric for the reward function in our methodology. The results, under both 1% and 10% training data configurations, exhibit a noticeable decline in performance without the lookahead component. This outcome unequivocally confirms the significance of these innovative elements in enhancing the method's effectiveness. Additionally, it was observed that ProDSP without the lookahead strategy (ProDSP (*w/o* user simulator LLM)) underperforms compared to ProDSP without user feedback (ProDSP (*w/o* user feedback LLM)) across the board. This discrepancy can be attributed to the fact that user feedback is integrated into the reward function with a specific weighting, whereas the lookahead heuristic plays a more pivotal role in the efficient generation of supportive responses.

5.2 Scenario 2: Bargain Negotiation

5.2.1 Experiment Setup

Dataset. In this scenario, our experiment is conducted on CraigslistBargain dataset (He et al., 2018). The dataset was created in a bargain negotiation setting, where the buyer and the seller negotiate the price of an item on sale, containing 11 negotiation strategies and 3466 cases.

Baseline. We compare several fine-tuned state-of-the-art (SOTA) baselines for negotiation dialogues, including **FeHED** (Zhou et al., 2019), and **DIALOGRAPH** (Joshi et al., 2021). In this task, we compare our method with two prompt-based LLM-enhanced method (with ChatGPT) **Proactive** and **ProCoT** proposed in (Deng et al., 2023b), which augments LLMs with the goal planning capability over descriptive reasoning chains.

Metrics. To evaluate the response generation, we employ **BLEU** and **BERTScore** as automatic metrics which is applied in (Deng et al., 2023b).We evaluate strategy prediction performance along with response generation quality, to assess strategy tracking. For strategy planning, we adopt **F1** and **AUC** for automatic evaluation on strategy planning.

Implementation. We also employ T5 (Raffel et al., 2020) as the fine-tuning model for negotiation strategy planning and leverage GPT-3.5-turbo (OpenAI, 2021) as the specific LLM which generates response and user simulation, which represents buyer and seller. GPT-4 (Achiam et al., 2023) is utilized as the AI feedback that provides user reward scores.

5.2.2 Experimental Results

Comparison with Baselines. The efficacy of our strategy planning approach on negotiation task is detailed in Table 2. We first compare the effectiveness of strategy planning and response generation ability with prompt-based LLM-enhanced method Proactive (Deng et al., 2023b) and ProCoT (Deng et al., 2023b). These two methods are claimed to be attampts of LLM-empowered methods for proactive dialogue systems by instructing LLMs with certain goal-oriented prompts. The experimental results in Table 2 has obviously demonstrated the difficulties for prompt-based models gaining planning and decision-making abilities, which also explaines the strength of our online RL framework with is conducted over a small fine-tuning policy model with the enhancement of frozen LLMs.

Besides, we also conduct comparision over several SOTA baselines in negotiation task, which incorporates 100% data during training procedures. As shown in Table 2, ProDSP has outperformed DIALOGRAPH on strategy planning F1 and AUC score with 2.4% and 0.6% respectively with only 10% training data involved, illustrating the low-resource demand and high efficiency of our proposed method. However, we noticed the decrease of fluency of the generated responses from ProDSP than FeHED and DIALOGRAPH. One reasonable explaination is the partation of training dataset involved for the LLMs to learn the expression from original corpus.

Ablation Study. In the ablation study on the negotiation task, we evaluated the effects of removing the long-term planning mode and the GPT-4-based reward collectors from our methodology.Based on both 1% and 10% training data configurations, reveal a significant drop in performance in the absence of the lookahead component. This result clearly underscores the importance of these innovative features in boosting the method's effectiveness. Moreover, it was found that ProDSP without online training (ProDSP (*w/o* user simulator LLM)) performs worse than ProDSP without user feedback (ProDSP (*w/o* user feedback LLM)) in all scenarios. This performance gap can be explained by the integration of user feedback into the reward function with a specific weighting, while the interactive setting is more crucial for the efficient generation of goal-oriented responses.

6 Conclusion

In conclusion, this paper introduces a proactive dialogue strategy planning (ProDSP) method designed to address the inherent limitations of existing systems. Our approach leverages a small, supervised fine-tuning language model to anticipate future strategy sequences, providing simulation hints that guide large language models (LLMs) in generating responses aligned with specific goals. This methodology is further refined through training within an interactive environment, utilizing an LLM-based user simulator to enhance the learning process.To evaluate online user feedback, we employ a GPT-4-based user simulator that quantifies goal-oriented rewards using multi-faceted metrics. This sophisticated feedback mechanism ensures that the responses generated by the model are both relevant and effective in achieving the desired outcomes. Through extensive experiments, we have demonstrated that our model surpasses competitive baselines in both strategy planning and dialogue generation tasks, particularly in scenarios requiring emotional support and negotiation.

Limitations

While our proposed method demonstrates competitive outcomes in the emotional support conversation and negotiation tasks, there are still deficiency about our proposed method. In our research, we leverage LLMs as a tool for generating responses, akin to a black-box utility, without delving into the potential enhancements achievable through fine-tuning with domain-specific expertise. This oversight suggests that incorporating expert knowledge in emotional support into the fine-tuning process of LLMs could yield even superior performance. Furthermore, the novel evaluate protocols should come along with the LLM-enhanced methods to replace the corpus-based evaluation metrics. However, this paper follows the main-stream methods to conduct comparison with SOTA approaches. Additionally, this paper studies two classic task of proactive dialogue, which is representative for the challenging strategy planning procedure, while the performance on other scenarios is uncertain.

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*.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. *arXiv preprint arXiv:2305.04147*.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empiri-*

cal Methods in Natural Language Processing, pages 3014–3026.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023a. A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750*.

Yang Deng, Wenqiang Lei, Lizi Liao, and Tat-Seng Chua. 2023b. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2023c. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023d. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *arXiv preprint arXiv:2305.10172*.

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*.

Adam M Grant and Susan J Ashford. 2008. The dynamics of proactivity at work. *Research in organizational behavior*, 28:3–34.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. *arXiv preprint arXiv:1808.09637*.

Zhiyuan Hu, Yue Feng, Yang Deng, Zekun Li, See-Kiong Ng, Anh Tuan Luu, and Bryan Hooi. 2023a. Enhancing large language model induced task-oriented dialogue systems through look-forward motivated goals. *arXiv preprint arXiv:2309.08949*.

Zhiyuan Hu, Yue Feng, Anh Tuan Luu, Bryan Hooi, and Aldo Lipani. 2023b. Unlocking the potential of user feedback: Leveraging large language model as user simulator to enhance dialogue system. *arXiv preprint arXiv:2306.09821*.

Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan Black, and Yulia Tsvetkov. 2021. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. *arXiv preprint arXiv:2106.00920*.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2020. End-to-end trainable non-collaborative dialog system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8293–8302.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding large language models via directional stimulus prompting. *arXiv preprint arXiv:2302.11520*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

OpenAI. 2021. Chatgpt: Openai's conversational ai.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. 2020. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. *arXiv preprint arXiv:2012.15375*.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2019. Alternating recurrent dialog model with large-scale pre-trained language models. *arXiv preprint arXiv:1910.03756*.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. *arXiv preprint arXiv:2305.17878*.

Yiheng Zhou, Yulia Tsvetkov, Alan W Black, and Zhou Yu. 2019. Augmenting non-collaborative dialog systems with explicit semantic and strategic dialog history. *arXiv preprint arXiv:1909.13425*.