Planning Like Human: A Dual-process Framework for Dialogue Planning

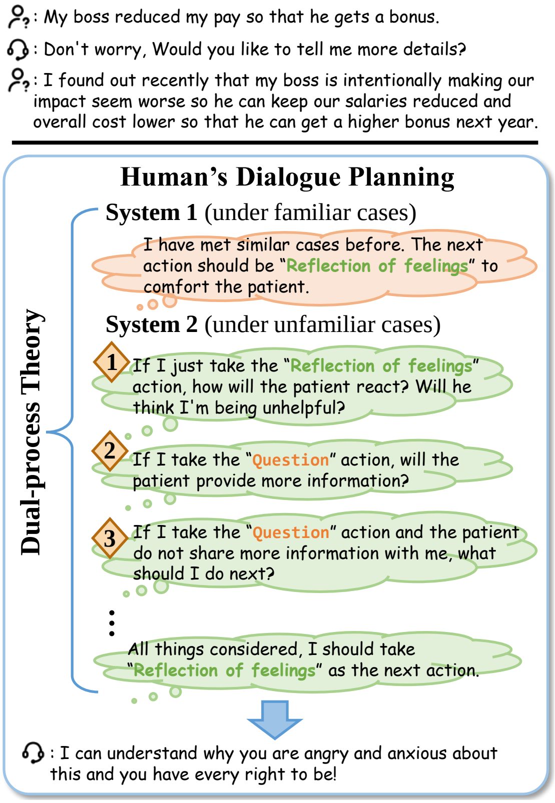
Tao He[1]*, **Lizi Liao**[2], **Yixin Cao**[3], **Yuanxing Liu**[1], **Ming Liu**[1,4]†, **Zerui Chen**[1], **Bing Qin**[1,4]
[1]Harbin Institute of Technology, Harbin, China
[2]Singapore Management University, Singapore
[3]School of Computer Science, Fudan University
[4]Peng Cheng Laboratory, Shenzhen, China
{the, yxliu, mliu, zrchen, qinb}@ir.hit.edu.cn,
lzliao@smu.edu.sg, caoyixin2011@gmail.com

Abstract

In proactive dialogue, the challenge lies not just in generating responses but in steering conversations toward predetermined goals, a task where Large Language Models (LLMs) typically struggle due to their reactive nature. Traditional approaches to enhance dialogue planning in LLMs, ranging from elaborate prompt engineering to the integration of policy networks, either face efficiency issues or deliver suboptimal performance. Inspired by the dual-process theory in psychology, which identifies two distinct modes of thinking—intuitive (fast) and analytical (slow), we propose the Dual-Process Dialogue Planning (DPDP) framework. DPDP embodies this theory through two complementary planning systems: an instinctive policy model for familiar contexts and a deliberative Monte Carlo Tree Search (MCTS) mechanism for complex, novel scenarios. This dual strategy is further coupled with a novel two-stage training regimen: offline Reinforcement Learning for robust initial policy model formation followed by MCTS-enhanced on-the-fly learning, which ensures a dynamic balance between efficiency and strategic depth. Our empirical evaluations across diverse dialogue tasks affirm DPDP's superiority in achieving both high-quality dialogues and operational efficiency, outpacing existing methods.[1]

1 Introduction

Large Language Models (LLMs), such as those described by (Ouyang et al., 2022; Touvron et al., 2023a,b), have revolutionized the field of natural language processing by demonstrating an unprecedented ability to understand context and generate coherent responses across a wide range of dialogue scenarios (Bang et al., 2023; Zhang et al., 2023b; Zhao et al., 2023). Despite these advances, LLMs predominantly operate in a reactive mode, often struggling to proactively guide conversations towards specific goals, a critical limitation for achieving truly dynamic interactions (Deng et al., 2023b). This gap underscores the urgent need for research into dialogue planning mechanisms that can strategically direct conversations, a topic that has been explored but remains an ongoing challenge (Levin et al., 1997; Liu et al., 2018; Cheng et al., 2022).


Human's Dialogue Planning
Dual-process Theory
System 1 (under familiar cases)
System 2 (under unfamiliar cases)
1
2
3

: My boss reduced my pay so that he gets a bonus.

: Don't worry, Would you like to tell me more details?

: I found out recently that my boss is intentionally making our impact seem worse so he can keep our salaries reduced and overall cost lower so that he can get a higher bonus next year.

I have met similar cases before. The next action should be "Reflection of feelings" to comfort the patient.

1 If I just take the "Reflection of feelings" action, how will the patient react? Will he think I'm being unhelpful?

2 If I take the "Question" action, will the patient provide more information?

3 If I take the "Question" action and the patient do not share more information with me, what should I do next?

⋮

All things considered, I should take "Reflection of feelings" as the next action.

: I can understand why you are angry and anxious about this and you have every right to be!

Figure 1: Using dual-process theory for dialogue planning in the human cognitive process. This is a case from ESConv (Liu et al., 2021). "Question" and "Reflection of feelings" are pre-defined dialogue actions in ESConv.

Dialogue planning, essential for shaping the trajectory of conversations to achieve desired outcomes, has seen various approaches (Zhang et al., 2020). Some studies aim to design more effective prompting procedures, such as Monte Carlo Tree Search (MCTS) (Väth et al., 2023;

---

*Work was done during an internship at SMU.

†Corresponding Author.

[1] Code available at: https://github.com/cs-holder/DPDP.git

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4768–4791
August 11-16, 2024 ©2024 Association for Computational Linguistics

Yu et al., 2023), resulting in outstanding performance achievements. However, these methodologies frequently encounter inefficiencies deemed unacceptable due to their complex and iterative nature. In contrast, studies exemplified by PPDPP (Deng et al., 2023b) employ Reinforcement Learning (RL) to train a pluggable small model as the policy network, circumventing the high-cost problem of optimizing LLMs and the low-efficiency issue of iterative prompting. However, in practical applications, the trained policy network predicts dialogue actions solely based on the current dialogue history while neglects potential user reactions in subsequent turns, hence resulting in limited performance.

Inspired by the dual-process theory of human cognition, as elaborated by Kahneman (Kahneman, 2003), which posits the existence of two distinct modes of thinking—System 1: fast, intuitive, and System 2: slow, analytical—we propose leveraging this framework for dialogue planning. Human conversationalists seamlessly integrate these systems, employing rapid, instinctual responses or engaging in deliberate, strategic thought as situations demand. This theory offers a compelling lens through which to reimagine dialogue planning, suggesting that a blend of intuitive and analytical planning could vastly enhance LLMs' ability to conduct proactive dialogues.

In response, we introduce the Dual-Process Dialogue Planning (DPDP) framework, a novel approach that incorporates two complementary planning systems: a neural policy LM model (System 1) for quick, instinctive responses to familiar situations, and an MCTS-based planner (System 2) for analytic, rational but slow planning in complex or novel scenarios. This framework allows for dynamic switching between systems based on policy LM's uncertainty, optimizing for both efficiency and depth of strategy. Key to the success of DPDP is the enhancement of the policy model's capability, which we address through a pioneering two-stage training approach. Initially, we employ offline RL to refine the policy model's base, mitigating the impact of suboptimal strategies and noise prevalent in training datasets. Subsequently, we leverage MCTS simulations to guide the policy model towards generating superior strategies, thereby accelerating its convergence and enhancing overall performance. Our comprehensive evaluation across various proactive dialogue tasks unequivocally demonstrates DPDP's superiority over contemporary methodologies, establishing new benchmarks in dialogue planning efficiency and efficacy.

In summary, our contributions are threefold:

- We present a dual-system approach to dialogue planning that mirrors human cognitive processes, balancing efficiency and strategic depth.
- We develop a novel two-stage training method for the policy model, integrating offline RL and MCTS to significantly enhance its performance.
- Experimental results across two datasets validate that our proposed framework effectively outperforms a series of baselines and performs more efficiently than MCTS-based methods.

2 Related Work

2.1 LLM-powered Dialogue Policy Planning

Dialogue planning, critical for guiding systems in task-oriented dialogues to achieve specific goals (Zhang et al., 2020), has been extensively explored (Jang et al., 2020; Takanobu et al., 2020; Jang et al., 2022; Feng et al., 2023; Wang et al., 2023). Despite advancements, the transition to leveraging Large Language Models (LLMs) introduces new challenges, primarily due to their static parameters and constrained capability for long-term planning (Yao et al., 2023; Xu et al., 2023). In response, several methods have been developed, ranging from intricate prompt engineering that encourages self-reflection (Deng et al., 2023a; Zhang et al., 2023a), to iterative feedback loops for planning enhancement (Fu et al., 2023), and even Monte Carlo Tree Search (MCTS) for identifying optimal actions (Yu et al., 2023). While these approaches achieve notable results, they are often marred by inefficiency and high operational costs. An alternative, presented by Deng et al. (2023b), is the Reinforcement Learning (RL)-based dialogue planning method PPDPP, which utilizes a compact model as a policy network for strategy prediction (Li et al., 2023), directly using dialogue history. Although this method offers efficiency, it falls short in simulating the nuanced human-like cognitive processes in proactive dialogue, particularly in anticipating the impact of actions on future dialogues. Our proposed solution draws from dual-process theory (Sun, 2002; Kahneman, 2003), marrying a policy LM planner for immediate strategy predictions with an MCTS planner for in-depth state simulations. This hybrid approach dynamically alternates between planners based on the system's confidence level, striking a balance between efficiency and strategic effectiveness.

2.2 Applications of Dual-process Theory

Existing studies have actively integrated System 1 and System 2 from dual-process theory (Kahneman, 2003) into machine learning methodologies. Mittal et al. (2017) view the vector space model and reasoning in knowledge graphs as fast and slow systems, respectively. Bengio (2017) highlights the close relationship between System 2 abilities and consciousness. Additionally, Chen et al. (2019) propose an end-to-end framework consisting of a generative decoder (fast thinking) and a reasoning module (slow thinking) to tackle complex tasks. Expanding on this, Liu et al. (2022) combine a neural network and a symbolic module by dual-process theory to address question-answering tasks. Inspired by above studies, we implement our dialogue planning framework with a fast policy LM planner and a slow MCTS planner.

2.3 Integrated Learning of RL and MCTS

In the field of RL, numerous investigations have converged on the amalgamation of MCTS and RL algorithms. Notably, some endeavors aim to leverage RL methodologies to enhance MCTS efficacy (Guo et al., 2016; Anthony et al., 2019; Soemers et al., 2019; Dieb et al., 2020). For instance, Guo et al. (2016) design reward-bonus functions to augment MCTS. Anthony et al. (2019) advocate the utilization of policy networks for the refinement of local policies and investigate planning without an explicit tree search. Concurrently, many studies adopt MCTS as a guiding mechanism for RL training, a paradigm commonly referred to as expert iteration (Efroni et al., 2018; Anthony et al., 2017; Grill et al., 2020). Prominent exemplars of this methodology encompass AlphaZero (Silver et al., 2017a,b) and MuZero (Schrittwieser et al., 2019). Inspired by AlphaZero, we apply the MCTS to enhance the policy LM in the dialogue planning task.

3 Methodology

3.1 Preliminaries

Problem formalization. Following existing work (Wang et al., 2020), we formulate the dialogue process as a Markov Decision Process (MDP). At each turn t, according to the observation on the current state s_t, i.e. the dialogue history $\{u_1^{sys}, u_1^{usr}, ..., u_t^{sys}, u_t^{usr}\}$, the dialogue system selects an action $a_t \in \mathcal{A}$, where u^{sys} and u^{usr} represent the system and user utterances, $\mathcal{A}$ is a set of candidate strategies pre-defined by domain experts. Then, guided by this action a_t, the system player generates the utterance u_{t+1}^{sys}. In return, the user player responds to the system with u_{t+1}^{usr}. This process repeats until the dialogue goal is achieved or the maximum number of turns T is reached. The objective is to learn a policy π_θ maximizing the expected cumulative rewards over observed dialogue episodes as:

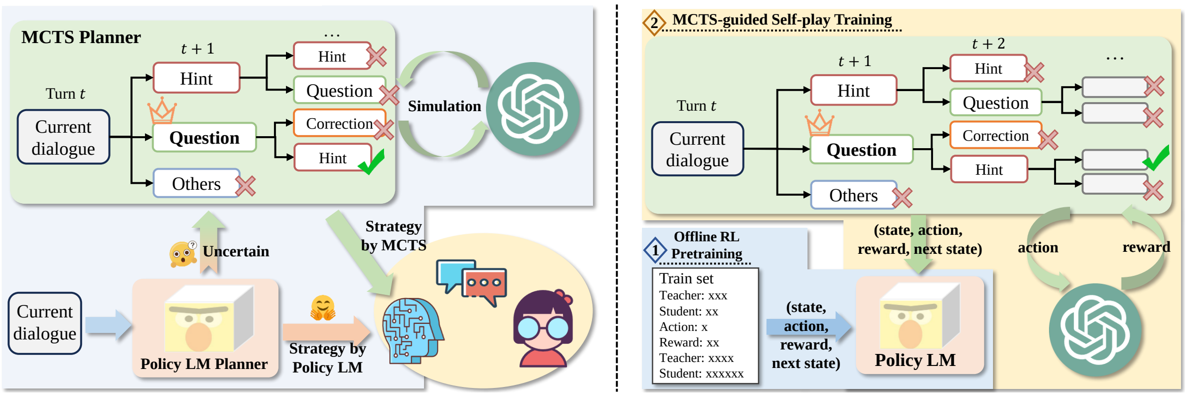
$$\pi^* = \arg\max_{\pi_\theta} \left[\sum_{t=0}^{T} r(s_t, a_t) \right], \tag{1}$$

where $r(\cdot)$ is a reward function, abbreviated as r_t.

LLM-powered role simulation. Following previous work (Deng et al., 2023b), we employ two LLMs as a user and an assistant to simulate dynamic user-assistant interactions. Role descriptions, along with instructions about their corresponding conversational goals, are delivered to each LLM. Furthermore, we prompt an LLM, following PPDPP (Deng et al., 2023b), to act as a critic for evaluating the dialogue states. The prompts utilized for role simulation and state evaluation processes remain consistent with those employed in PPDPP. Please refer to Appendix G for detailed prompts. Through this approach, we can concentrate our research efforts on effectively planning strategies for each dialogue turn.

3.2 Dual-process Planning Framework

We present the dual-process dialogue planning framework in Figure 2(a). Motivated by human cognitive research, human cognition and behavior are propelled by two cognitive systems: the intuitive and analytic systems. In our study, a smaller model is trained to function as an intuitive policy LM (we implement using RoBERTa-large), capable of directly predicting the next conversational action based on the dialogue history. MCTS is applied as an analytic process by iteratively simulating subsequent dialogue turns to select an approximately optimal strategy through multiple simulations. If the policy LM is not confident enough in the current state, we shift to employing MCTS for action planning. We propose a nonparameterized control gate mechanism for deciding the switch.



(a) Overall Procedure of DPDP

(b) Two-stage Training for Policy LM

Figure 2: DPDP combines two planners: (a) a Policy LM for quick responses and an MCTS planner for complex scenarios, switching when Policy LM is uncertain during inference. (b) We propose a two-stage training approach for Policy LM, involving pretraining over static data with offline RL and further finetuning with MCTS simulations.

3.2.1 Policy LM Planner

We propose utilizing a tunable pre-trained language model, e.g., RoBERTa (Liu et al., 2019), as the dialogue policy planner to control the dialogue process. Different from previous methods (Deng et al., 2023b), we involve not only a policy network but also a Q-network. Our design rationale is based on two points: (1) The CIMA training set comprises only dialogue snippets rather than complete dialogue histories. It is inadequate to only learn a policy network in our proposed Offline RL-based pretraining method. An action-value function, i.e., Q-Network, is required to aid in training the policy network. (2) We utilize the LLM as the reward function, that is we view the critic LLM as part of the environment. Modeling the environment is a typical method to reduce interaction while maintaining performance (Luo et al., 2022).

By connecting two different MLP layers, the policy LM planner comprises a policy network $\pi_\theta(a|s)$ for action prediction and a Q-network $Q_\beta(a|s)$ for state evaluation. At each turn, the subsequent strategy a_t is predicted by inputting the dialogue state s_t into the policy network.

3.2.2 MCTS Planner

Following GDPZero (Yu et al., 2023), we utilize MCTS (Weber, 2010; Liebana et al., 2015) for simulating subsequent strategic deliberation, a process typically involving four stages: Selection, Expansion, Evaluation, and Backpropagation. After multiple simulations, the conversational strategy for the next turn of response is determined based on the action most frequently applied during these processes. For further details about MCTS in dialogue planning, please refer to Appendix D.

Different from GDPZero, we apply policy LM to produce prior knowledge for MCTS. To establish the prior action distribution for each Selection step, GDPZero computes action probabilities by repeatedly prompting an LLM. Conversely, we utilize the domain-specific trained policy LM to generate the prior probability. This approach not only enables the application of learned domain-specific knowledge for improved initialization but also diminishes the frequency of calls to the LLM, thus enhancing efficiency and reducing costs. We verify this operation in the following experiments.

3.2.3 Synergizing Two Planners

Our objective is to synergistically utilize both the policy LM planner and MCTS planner to establish an adaptive dual-process system. Similar to human cognitive processes during proactive dialogue, individuals can rapidly respond to appropriate strategies when encountering familiar conversational states. Conversely, when encountering unfamiliar states, it becomes necessary to select the most suitable strategy by simulating potential reactions in subsequent dialogue turns. In our framework, we give priority to utilizing a policy LM for action selection. If the policy LM detects inadequate confidence regarding the current state, we switch to employing MCTS for action planning.

To reduce reliance on training, we propose a non-parameterized control gate mechanism to control switching. For the action distribution $\pi_\theta(a_t|s_t)$ predicted by policy LM, we assess the uncertainty by the probability difference for the top-2 values: $\delta(\pi_\theta(a_t|s_t)) = top(1) - top(2)$, where $top(i)$

means the i-th largest value in $\pi_\theta(a_t|s_t)$. If this difference surpasses a threshold η, it indicates that the policy LM has high confidence in the current decision-making, prompting the utilization of the policy LM for action selection. Conversely, small differences imply that the policy LM is uncertain, prompting the utilization of MCTS for action selection. By setting appropriate η, we can roughly control the ratio of MCTS used. For details on how to determine η, please see Appendix B. Of course, other uncertainty measures like entropy can also be applied in our framework.

3.3 Two-stage Training for Policy LM

3.3.1 Offline RL-based Pretraining

Our two-stage training scheme for the Policy LM, depicted in Figure 2(b), begins with offline RL-based Pretraining to ensure effective initialization, aiming to reduce interaction time during subsequent online learning. Unlike direct supervised learning, which can introduce biases through suboptimal or noisy data, offline RL (Kumar et al., 2020) refines pretraining by using soft rewards to discern valuable strategies from the dataset, avoiding the pitfalls of hard labels. In detail, our approach initially assigns scores to each dialogue turn in the training set using the critic LLM, which serves as rewards for annotated dialogue strategies. As a result, we reconstruct an MDP corpus comprising complete states, actions, and rewards. Utilizing this corpus, we pretrain the policy LM, which consists of the policy network $\pi_\theta(a|s)$ and the Q-network $Q_\beta(s,a)$.

Specifically, the pretraining details for ESConv (Liu et al., 2021) and CIMA (Stasaski et al., 2020) datasets differ due to the absence of complete dialogue trajectories on CIMA. We outline the optimization process for ESConv here, while details for CIMA are provided in Appendix A. For the policy network, we optimize it by:

$$\mathcal{L}_{pre,\theta} = -\sum_{t=1}^{T} \hat{Q}(s_t,a_t)\log \pi_\theta(a_t|s_t), \quad (2)$$

where $\hat{Q}(s_t,a_t)=\sum_{t=1}^{T}\gamma^t R(a_t|s_t)$ represents cumulative rewards, γ means the discount factor, and $R(a_t|s_t)$ is the received reward by selecting action a_t upon the state s_t. We use the same strategy as PPDPP, which generates 10 evaluations on the current state, maps each evaluation into a predefined score, and finally computes the mean value as the reward (Deng et al., 2023b). For ESConv, we map “feel worse”, “feel the same”, “feel better”, “solved” into -1.0, -0.5, 0.1, 1.0. For the Q-network $Q_\beta(s,a)$, we optimize it to approximate $\hat{Q}(s,a)$:

$$\mathcal{L}_{pre,\beta} = \sum_{t=1}^{T} \mathrm{MSE}(Q_\beta(s_t,a_t),\hat{Q}(s_t,a_t)). \quad (3)$$

Finally, the overall optimizing loss for the pretraining stage is:

$$\mathcal{L}_{pre} = \mathcal{L}_{pre,\theta} + \lambda_1 * \mathcal{L}_{pre,\beta}, \quad (4)$$

where λ_1 is a hyperparameter to control loss weight. By doing this, we expect to learn a better initialization compared to direct supervised learning.

3.3.2 MCTS-guided Self-play Training

Additional interaction with the environment is necessary since static training sets cannot cover the entire state-action space. In interactive online learning, we initiate two LLMs to simulate self-play dialogues between the user and the assistant. Given the current state s_t, rather than directly utilizing the policy agent to predict the next action, we utilize MCTS for action prediction. The predicted action is then mapped to a pre-defined natural language instruction, $\mathcal{M}_a(a_t)$. Subsequently, the dialogue history s_t along with $\mathcal{M}_a(a_t)$ will trigger the LLM to generate the appropriate system response, then prompt the LLM to generate the corresponding user response. Following this, the dialogue process transitions to a new state s_{t+1}. We employ an LLM as a critic to calculate the action reward r_t. The collected transition records $\{s_t, a_t, s_{t+1}, r_t\}$ are used to train the policy model.

We optimize the policy LM using the Actor-Critic algorithm instead of the REINFORCE algorithm (Sutton et al., 1999). The optimization loss for the Q-network is as follows:

$$\mathcal{L}_{sp,\beta} = \sum_{t=1}^{T}[Q^*(s_t,a_t) - Q_\beta(s_t,a_t)]^2,$$
$$Q^*(s_t,a_t) = R(a_t|s_t) + \gamma * \max_{a'} Q_\beta(s_{t+1},a'),$$

and for the policy network as:

$$\mathcal{L}_{sp,\theta} = \sum_{t=1}^{T}[(Q_\beta(s_t,a_t) - \hat{Q}(s_t,a_t)) * \log \pi_\theta(a_t|s_t)].$$

where Q_β is the Q-network to calculate state-action values. $\hat{Q}(s_t,a_t)$ is cumulative rewards defined before. Finally, the overall optimizing loss for the self-play training phase is:

$$\mathcal{L}_{sp} = \mathcal{L}_{sp,\theta} + \lambda_2 * \mathcal{L}_{sp,\beta}, \tag{5}$$

where λ_2 is also a loss weight.

4 Experiments

4.1 Datasets

We evaluate the proposed framework on three proactive dialogue datasets, including ESConv (Liu et al., 2021) (emotional support dialogue), CIMA (Stasaski et al., 2020) (tutoring dialogue), and CraigslistBargain (or CB, negotiating prices) (He et al., 2018). ESConv is split into 1040/130/130 cases for training/valid/test set, with pre-defined 8 actions. CIMA is split into 909/113/113 cases for training/valid/test set, with 5 dialogue actions. ESConv and CIMA are collaborative dialogue tasks where both participants share the same goal. In contrast, CB is a non-collaborative dialogue task where the buyer aims for the lowest price, and the seller aims for the highest. CB consists of 3290 training cases, 188 valid cases, and 188 testing cases, involving 11 buyer bargaining actions. Please refer to Appendix G.4 for pre-defined dialogue actions. Following the settings of PPDPP (Deng et al., 2023b), we use human-annotated dialogues in the train set for pretraining. For the self-play training phase, we only use the case background information in the dataset for state initialization.

4.2 Baselines

We aim to demonstrate the efficacy of this framework by primarily comparing against PPDPP. Additionally, we follow PPDPP to compare with a general fine-tuning dialogue model DialoGPT (Zhang et al., 2019), and a range of prompting-based methods, including Standard Prompting, Proactive (Deng et al., 2023a), ProCoT (Deng et al., 2023a), Ask-an-Expert (Zhang et al., 2023a), and ICL-AIF (Fu et al., 2023). Following Deng et al. (2023b), we report the results of baselines.

4.3 Evaluation Metrics

For automatic evaluation, we employ two key metrics: the average turn (AT) and the success rate (SR). AT measures goal completion efficiency by calculating the average number of turns required to achieve the goal, while SR measures goal completion effectiveness by computing the success rate of achieving the goal within a predefined maximum number of turns. For CB, following PPDPP, we also use SL (Sale-to-List Ratio) to evaluate the buyer's deal. A higher SL represents the buyer gets more benefits from deals. If the deal fails, we assign SL as 0. Additionally, upon analyzing specific examples, we discovered biases in directly using ChatGPT for evaluation in ESConv. Therefore, we also conduct human evaluation for comparison. Three annotators compare the generated responses from four perspectives: Suggestion **Sug.**, Identification **Ide.**, Comforting **Com.**, and Overall **Ove.**. The instruction for each perspective is presented in Appendix E. Each annotator is asked to determine whether DPDP (policy LM) outperforms PPDPP, with possible answers being win, lose, or tie for these four aspects. Finally, we average the results from three annotators.

4.4 Experimental Details

In line with PPDPP (Deng et al., 2023b), we instantiate the policy LM utilizing RoBERTa-large (Liu et al., 2019). We implement MCTS using codes from GDPZero (Yu et al., 2023). We also employ gpt-3.5-turbo-0613 as the static LLM for both the playing system and the user, alongside the reward model. All used prompts and temperatures are consistent with PPDPP. Additionally, we adhere to the same mapping $\mathcal{M}_a$ of natural language instructions with PPDPP for dialogue actions. For more training details, please refer to the Appendix C.

4.5 Results and Analysis

4.5.1 Overall Performance

Automatic evaluation results. Table 1 & 2 summarizes the experimental results of method comparisons on three datasets. On ESConv and CIMA, the proposed DPDP method, integrating both System 1 and System 2, consistently outperforms all the baselines by a noticeable margin. We further analyze the influence of different training methods and systems. Initially, we demonstrate the effectiveness of using System 1 (or Policy LM) alone. Compared to PPDPP, System 1 also utilizes only the policy network to predict subsequent actions. System 1, trained only through pretraining or self-play training, consistently outperformed PPDPP, affirming the efficacy of both offline RL-based pretraining and MCTS-guided self-play training. Furthermore, combining both training methods in System 1 yields further improvements, validating the

Models	ESConv		CIMA	
	AT↓	SR↑	AT↓	SR↑
DialoGPT (Zhang et al., 2019)	5.31	0.7538	5.43	0.4956
Standard	5.10	0.7692	3.89	0.6903
AnE (Zhang et al., 2023a)	4.76	0.8000	3.86	0.6549
Proactive (Deng et al., 2023a)	5.08	0.7538	4.84	0.5310
ProCoT (Deng et al., 2023a)	4.75	0.7923	4.58	0.5487
ICL-AIF (Fu et al., 2023)	4.69	0.8079	4.19	0.6106
PPDPP (Deng et al., 2023b)	4.56	0.8462	3.03	0.8407
DPDP (System 1)	3.61	0.9000	**2.24**	0.9469
-w/o PT	4.22	0.8769	2.36	0.9292
-w/o SPT	3.97	0.8692	2.51	0.8938
DPDP (System 2)	2.13	0.9923	2.49	0.9735
DPDP (System 1&2)	**2.13**	**0.9923**	2.28	**0.9823**

Table 1: Experimental results on ESConv and CIMA. PT means Pretraining and SPT means Self-Play Training.

Models	CraisglistBargain		
	AT↓	SR↑	SL↑
DialoGPT (Zhang et al., 2019)	6.73	0.3245	0.2012
Standard	6.47	0.3830	0.1588
AnE (Zhang et al., 2023a)	5.91	0.4521	0.2608
Proactive (Deng et al., 2023a)	5.80	0.5638	0.2489
ProCoT (Deng et al., 2023a)	6.22	0.5319	0.2486
ICL-AIF (Fu et al., 2023)	6.53	0.3617	0.1881
PPDPP (Deng et al., 2023b)	5.62	0.6117	0.3376
-w/o SFT	5.71	0.6223	0.3354
-w/o RL	5.57	0.6649	0.2280
DPDP (System 1)	5.03	0.7447	**0.4108**
DPDP (System 1&2, 22.3% MCTS)	3.69	0.8298	0.3102
DPDP (System 1&2, 51.4% MCTS)	2.77	0.9468	0.3118
DPDP (System 1&2, 60.3% MCTS)	**2.49**	0.9681	0.2856
DPDP (System 2)	2.78	**0.9734**	0.2728

Table 2: Experimental results on CraisglistBargain.

rationality of the two-stage training approach. System 2, with multiple simulations, achieved superior results compared to System 1, underscoring the effectiveness of MCTS. However, results on CIMA demonstrate that relying solely on MCTS may not lead to optimal outcomes. By appropriately combining System 1 and System 2, better results can be achieved while maintaining higher efficiency (requiring fewer dialogue turns to complete).

For CraigslistBargain, firstly, compared to the previous state-of-the-art method, DPDP based on System 1 (Policy LM) significantly improves AT (5.57→ 5.03), SR (0.6649→ 0.7447), and SL (0.3376→ 0.4108), demonstrating that our two-stage training method not only enhances deal success rates but also increases benefits. Secondly, System 2 (MCTS) significantly increases the success rates but reduces benefits. This suggests that the System 2 is more prone to compromise. The reason may be that System 2 doesn't require training and terminates the dialogue once the benefits exceed the pre-defined threshold; while the optimized Policy LM focuses on higher-benefits dialogues from MCTS, thereby improving benefits compared with previous work while maintaining a high deal success rate. These results also indicate that DPDP needs further improvement for non-collaborative proactive dialogue tasks in the future.

Human evaluation results. Building upon prior research (Liu et al., 2021; Joshi et al., 2021), we conduct human evaluations on 50 randomly selected dialogues from ESConv. We focus on ESConv due to its pronounced automatic evaluation bias, primarily stemming from the subjective nature of assessing a patient's state. We present the results in Figure 3. We also provide a human evaluation of CIMA in Appendix F. It is evident that our approach differs significantly from PPDPP in the first two criteria: DPDP (System 1) tends to provide advice, whereas PPDPP leans towards expressing empathy, with comparable proficiency in understanding the patient's condition. Given that offering practical advice might be more beneficial than merely expressing empathy in actual psychological counseling, this disparity elucidates why DPDP (System 1) scores considerably higher than PPDPP in the Overall aspect.

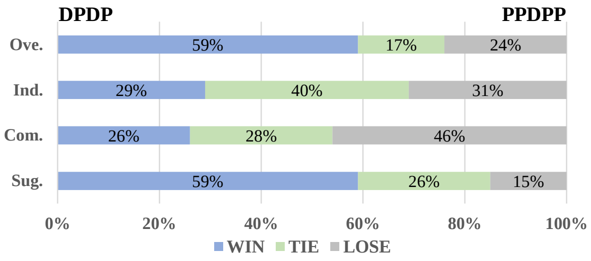


Figure 3: Human evaluation results on ESConv.

4.5.2 Trade-off between Two Planners

Our framework integrates the capabilities of both policy LM and MCTS planners, enabling seamless dynamic transitions between these methods during operation. To assess the impact of MCTS, we explored variations in Success Rate (SR) and Average Turns (AT) under differing degrees of MCTS engagement. The results are detailed in Table 3.

For the ESConv dataset, our analysis revealed a marked increase in SR and a significant decrease in AT as the involvement of MCTS escalated during

MCTS Ratios	**ESConv** AT↓	SR↑	**MCTS Ratios**	**CIMA** AT↓	SR↑
0.0%	3.61	0.9000	**0.0%**	2.24	0.9469
21.9%	3.42	0.9154	**28.6%**	2.39	0.9646
46.5%	2.95	0.9692	**50.0%**	2.28	0.9823
68.3%	2.72	0.9769	**81.1%**	2.58	0.9735
100.0%	2.13	0.9923	**100.0%**	2.49	0.9735

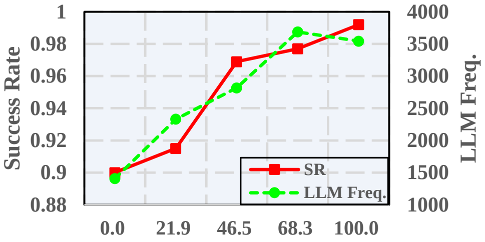
Table 3: SR and AT results of employing different application ratios of MCTS on ESConv and CIMA.

inference. This indicates a clear benefit of incorporating MCTS in more complex dialogue scenarios where strategic planning is crucial. Conversely, the results for the CIMA dataset painted a different picture. Here, we observed an improvement in performance metrics up to a 50% MCTS involvement threshold, beyond which the benefits diminished, eventually leading to a decline in performance. This pattern suggests that for tasks requiring specific reactions, such as providing hints or correcting translations, excessive reliance on MCTS for long-term planning is not only unnecessary but can be detrimental. This phenomenon, which we describe as *overthinking leads to unforeseen mistakes in simpler tasks*, echoes findings from (Ma et al., 2023), highlighting the potential pitfalls of over-reliance on extensive simulations for tasks that demand immediate and straightforward responses.

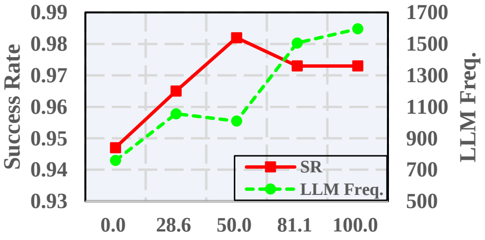
The optimal performance with balanced MCTS involvement showcases our framework's ability to blend planning strengths, achieving efficiency and effectiveness tailored to task demands, demonstrating its adaptability and validating its design.

4.5.3 Cost & Efficiency Analysis

MCTS typically enhances performance but suffers from elevating the frequency of invoking the LLM (i.e., ChatGPT). Selecting an action that will be actually performed with policy LM only needs 3 LLM calls (1 call for generating system utterance, 1 call for generating user utterance, and 1 for critic), whereas MCTS may require up to 3*10 calls, where 10 is the MCTS simulation times for determining one action. Given that the majority of the inference phase is spent awaiting LLM responses, we assess efficiency and cost based on the frequency of LLM invocations. We analyze the evolving trends in the frequency of LLM's calls across various MCTS participation ratios, as depicted in Figure 4. The results reveal that as the utilization of MCTS increases, the frequency of LLM calls gradually escalates, resulting in a heightened application cost.



(a) ESConv



(b) CIMA

Figure 4: Curves depicting the variation in LLM usage frequency and success rate on ESConv and CIMA as the proportion of MCTS is used.

Nevertheless, anomalies also occur at x=100% in ESConv and x=50.0% in CIMA. This reason is that while MCTS usage increases, enhancements in system capabilities lead to a reduction in the average turn. Consequently, the used times of MCTS increase while the LLM invocation count may decrease. By selecting a suitable MCTS ratio, we can strike a balance between effectiveness and the frequency of LLM usage.

4.5.4 Influence of Policy LM on MCTS

Note that we utilize the policy LM to produce prior probabilities for MCTS simulations. Hence, we further examine the influence of the policy LM on MCTS planning here. Specifically, we employ policy LMs trained via pretraining (PT), self-play training (SPT), and both pretraining and self-play (PT& SPT), respectively, to supply prior probabilities to MCTS. The whole evaluation solely relies on the MCTS planner. Additionally, we present results obtained by directly utilizing the uniform distribution to initialize prior probabilities and prompting an LLM to calculate the prior probabilities like GDPZero (Yu et al., 2023) for comparison. The results are shown in Table 4.

The results indicate that the performance of the policy LM indeed impacts MCTS, particularly when the policy LM's performance is substantially improved. On ESConv, although the improvements are minor, when the policy LM performs optimally, MCTS also demonstrates the best performance. On

(a) ESConv

Planner	PT	SPT	PT&SPT	Unif.	Chat.
Policy LM	0.8692	0.8769	0.9000	-	-
MCTS	0.9769	0.9692	0.9923	0.9846	0.9769

(b) CIMA

Planner	PT	SPT	PT&SPT	Unif.	Chat.
Policy LM	0.8938	0.9292	0.9469	-	-
MCTS	0.9469	0.9823	0.9735	0.7699	0.7522

Table 4: SR results using MCTS initialized with different prior distributions on ESConv and CIMA. **Unif.** indicates uniform distribution, while **Chat.** means using ChatGPT for initialization like in GDPZero.

CIMA, the sharp decline at SR when initializing priors with uniform distribution and the GDPZero method further underscores the significance of effective prior probabilities and the critical role of utilizing the policy LM to furnish prior knowledge. Concurrently, from another perspective, it also illustrates that our self-play training method is an iterative ascending process. As the training progresses, the capability of policy LM improves, providing better prior knowledge and enhancing the planning ability of MCTS, which in turn guides policy LM to further improvement. The performance change significantly varies on ESConv and CIMA when using a uniform distribution and GDPZero. On ESConv, therapists often have multiple valid strategy options for the same dialogue state, i.e., providing suggestions or expressing empathy for the same person faced with job crisis are both acceptable. Conversely, on CIMA, the suitable strategy is often limited based on the student's translation state (e.g. Hint for student's question, and Confirmation to confirm the student's answer). Therefore, it is more challenging to find the unique valid action by MCTS and thus relying on prior knowledge becomes more significant on CIMA.

5 Conclusion and Future Work

In this work, we introduced a dialogue planning framework that leverages the dual-process theory of human cognition, strategically alternating between quick, intuitive responses and detailed, analytical planning to emulate human-like conversational dynamics. To bolster this framework's capabilities, we implemented a two-stage training strategy, merging offline reinforcement learning for foundational training with advanced MCTS-guided self-play for refinement. The resulting empirical evidence demonstrated our dual-process framework's superior performance against leading methods, showcasing significant advancements in dialogue planning. Moving forward, our research will aim to refine the switching mechanism between planning modes and further optimize MCTS use, reducing computational expenses and enhancing dialogue planning towards achieving more nuanced, human-like interactions.

Limitations

Evaluation Quality. DPDP prompts an LLM (e.g. ChatGPT) to perform dialogue simulation and value estimation. Although LLMs have been applied to various tasks in data quality or data value assessment (Stiennon et al., 2020; Bai et al., 2022; Gilardi et al., 2023; He et al., 2023), demonstrating good evaluation performance and efficient evaluation costs, we found significant evaluation bias in our experiments. For instance, misjudging the patient's state led to premature termination of treatment. The average dialogue turns in DPDP did not exceed 3, which is clearly unrealistic: it is unlikely to resolve the patient's issues in just two dialogue turns. Similarly, PPDPP also faces similar issues. The problem of evaluation bias not only affects the final metric calculations but also influences the obtained rewards during the training process. To mitigate this issue, we conducted human evaluation, but the high cost of manual assessment impacts its large-scale use, thus preventing the correction of evaluation bias during the training phase.

Optimization cost. Our approach differs from previous prompt-based methods in that it requires training, especially in the second stage of self-play training, which involves continuous interaction with LLMs. The problem exacerbates due to the necessity of employing MCTS for multiple simulations per dialogue turn. Despite limiting the training to only 5 and 3 epochs respectively on ESConv and CIMA, aided by MCTS guidance, as opposed to the 10 epochs in PPDPP training, there remains a significant increase in ChatGPT API call. We suspect that further reduction in training costs can be achieved by enhancing the utilization of MCTS interaction history, such as employing all interaction records for following MCTS simulations, rather than only focusing on the most optimal actions chosen per turn.

Ethics Statement

Our work presents an algorithm for dialogue policy planning aimed at enhancing the development of future dialogue systems and improving their effectiveness in assisting users or systems in accomplishing tasks and goals. Generally, while most algorithms are not designed for unethical usage, there is often potential for abuse in their applications. In our experiments using ESConv (Liu et al., 2021) and CIMA (Stasaski et al., 2020), we employ DPDP to facilitate emotionally supportive conversations with patients and to tutor students in translation tasks. Nevertheless, due to DPDP's inherent goal-agnostic nature, there exists the potential for its unethical use, including fraudulent activities like scamming. We explicitly reject any employment of DPDP for unlawful or morally unjust endeavors.

Acknowledgements

This research was supported by the National Key Research and Development Project (2022YFF0903301), the National Science Foundation of China (U22B2059, 62276083), Shenzhen Foundational Research Funding (JCYJ20200109113441941), Major Key Project of PCL (PCL2021A06). This work was also supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 23-SIS-SMU-010). We are sincerely grateful to all reviewers for their insightful feedback.

References

Thomas W. Anthony, Robert Nishihara, Philipp Moritz, Tim Salimans, and John Schulman. 2019. Policy gradient search: Online planning and expert iteration without search trees. *ArXiv*, abs/1904.03646.

Thomas W. Anthony, Zheng Tian, and David Barber. 2017. Thinking fast and slow with deep learning and tree search. In *Neural Information Processing Systems*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Yoshua Bengio. 2017. The consciousness prior. *ArXiv*, abs/1709.08568.

Di Chen, Yiwei Bai, Wenting Zhao, Sebastian Ament, J. Gregoire, and Carla Pedro Gomes. 2019. Deep reasoning networks: Thinking fast and slow. *ArXiv*, abs/1906.00855.

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 *Conference on Empirical Methods in Natural Language Processing*.

Yang Deng, Wenqiang Lei, Lizi Liao, and Tat-Seng Chua. 2023a. 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. 2023b. Plug-and-play policy planner for large language model powered dialogue agents. *arXiv preprint arXiv:2311.00262*.

Sae Dieb, Zhilong Song, Wan-Jian Yin, and Masashi Ishii. 2020. Optimization of depth-graded multilayer structure for x-ray optics using machine learning. *Journal of Applied Physics*, 128:074901.

Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. 2018. How to combine tree-search methods in reinforcement learning. *ArXiv*, abs/1809.01843.

Yihao Feng, Shentao Yang, Shujian Zhang, Jianguo Zhang, Caiming Xiong, Mingyuan Zhou, and Huan Wang. 2023. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems. *arXiv preprint arXiv:2302.10342*.

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*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Jean-Bastien Grill, Florent Altch'e, Yunhao Tang, Thomas Hubert, Michal Valko, Ioannis Antonoglou, and Rémi Munos. 2020. Monte-carlo tree search as regularized policy optimization. *ArXiv*, abs/2007.12509.

Xiaoxiao Guo, Satinder Singh, Richard L. Lewis, and Honglak Lee. 2016. Deep learning for reward design to improve monte carlo tree search in atari games. *ArXiv*, abs/1604.07095.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. *ArXiv*, abs/1808.09637.

Xingwei He, Zheng-Wen Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. Annollm: Making large language models to be better crowdsourced annotators. *ArXiv*, abs/2303.16854.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2020. Bayes-adaptive monte-carlo planning and learning for goal-oriented dialogues. In *AAAI Conference on Artificial Intelligence*.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2022. Gpt-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems. In *10th International Conference on Learning Representations, ICLR 2022*. International Conference on Learning Representations, ICLR.

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

Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93:1449–1475.

Aviral Kumar, Aurick Zhou, G. Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *ArXiv*, abs/2006.04779.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 1997. Learning dialogue strategies within the markov decision process framework. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 72–79.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xi Yan. 2023. Guiding large language models via directional stimulus prompting. *ArXiv*, abs/2302.11520.

Diego Perez Liebana, Jens Dieskau, Martin Hunermund, Sanaz Mostaghim, and Simon M. M. Lucas. 2015. Open loop search for general video game playing. *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*.

Bing Liu, Gökhan Tür, Dilek Z. Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *North American Chapter of the Association for Computational Linguistics*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *ArXiv*, abs/2106.01144.

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

Zhixuan Liu, Zihao Wang, Yuan-Chun Lin, and Han Li. 2022. A neural-symbolic approach to natural language understanding. *ArXiv*, abs/2203.10557.

Fan Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. 2022. A survey on model-based reinforcement learning. *ArXiv*, abs/2206.09328.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Conference on Empirical Methods in Natural Language Processing*.

Sudip Mittal, Anupam Joshi, and Timothy W. Finin. 2017. Thinking, fast and slow: Combining vector spaces and knowledge graphs. *ArXiv*, abs/1708.03310.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, L. Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy P. Lillicrap, and David Silver. 2019. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604 – 609.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, L. Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2017a. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *ArXiv*, abs/1712.01815.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017b. Mastering the game of go without human knowledge. *Nature*, 550:354–359.

Dennis J. N. J. Soemers, Éric Piette, Matthew Stephenson, and Cameron Browne. 2019. Learning policies from self-play with policy gradients and mcts value estimates. *2019 IEEE Conference on Games (CoG)*, pages 1–8.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. Cima: A large open access dialogue dataset for tutoring. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Nisan Stiennon, Ouyang Long, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Francis Christiano. 2020. Learning to summarize with human feedback. In *Neural Information Processing Systems*.

Ron Sun. 2002. Duality of the mind.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Y. Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Neural Information Processing Systems*.

Ryuichi Takanobu, Runze Liang, and Minlie Huang. 2020. Multi-agent task-oriented dialog policy learning with role-aware reward decomposition. *ArXiv*, abs/2004.03809.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Dirk Väth, Lindsey Vanderlyn, and Ngoc Thang Vu. 2023. Conversational tree search: A new hybrid dialog task. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Jian Wang, Dongding Lin, and Wenjie Li. 2023. Dialogue planning via brownian bridge stochastic process for goal-directed proactive dialogue. *ArXiv*, abs/2305.05290.

Sihan Wang, Kaijie Zhou, Kunfeng Lai, and Jianping Shen. 2020. Task-completion dialogue policy learning via monte carlo tree search with dueling network. In *Conference on Empirical Methods in Natural Language Processing*.

Richard Weber. 2010. Optimization and control. *University of Cambridge*.

Canwen Xu, Yichong Xu, Shuo Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023. Small models are valuable plug-ins for large language models. *ArXiv*, abs/2305.08848.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phí Th Mùi, Haiquan Wang, Caiming Xiong, and Silvio Savarese. 2023. Retroformer: Retrospective large language agents with policy gradient optimization. *ArXiv*, abs/2308.02151.

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

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

Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen M. Meng. 2023b. Sgp-tod: Building task bots effortlessly via schema-guided llm prompting. *ArXiv*, abs/2305.09067.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zheng Zhang, Lizi Liao, Xiaoyan Zhu, Tat-Seng Chua, Zitao Liu, Yi-Feng Huang, and Minlie Huang. 2020. Learning goal-oriented dialogue policy with opposite agent awareness. *ArXiv*, abs/2004.09731.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *ArXiv*, abs/2304.09582.

A Pretraining for CIMA

Due to the absence of complete trajectories in the training set of CIMA, only dialogue fragments are available. Therefore, we can not apply the same optimization method as on ESConv. In this section, we introduce the pretraining process on CIMA.

For the policy network, since actual state-action values cannot be obtained without complete trajectories, we utilize a Q-Network for approximation. The optimization loss is as follows:

$$\mathcal{L}_{pre,\theta} = \sum_{t=1}^{T} [(\hat{Q}_\beta(s_t, a_t) - Q^*(s_t, a_t)) \log \pi_\theta(a_t|s_t)],$$
$$Q^*(s_t, a_t) = R(a_t|s_t) + \gamma * \max_{a'} Q_\beta(s_{t+1}, a'),$$

where $\hat{Q}_\beta(s_t, a_t)$ means using the Q-network Q_β to calculate state-action values with stopping gradient. $R(a_t|s_t)$ is the reward calculated by the same method as on ESConv but with different maps. We map "incorrect answer", "no answer", "partially correct answer" and "correct answer" as -1.0, -0.5, 0.5, and 1.0. Simultaneously, the optimization loss for Q-network is as follows:

$$\mathcal{L}_{pre,\beta} = \sum_{t=1}^{T} [Q^*(s_t, a_t) - Q_\beta(s_t, a_t)]^2,$$
$$Q^*(s_t, a_t) = R(a_t|s_t) + \gamma * \max_{a'} Q_\beta(s_{t+1}, a').$$

Finally, the overall loss for pretraining on CIMA is as follows:

$$\mathcal{L}_{pre} = \mathcal{L}_{pre,\theta} + \lambda_1 * \mathcal{L}_{pre,\beta}. \quad (6)$$

B Method to Determine the MCTS Ratio

In practical applications, while it is hard to precisely control the proportion of MCTS usage, we can approximate it in an easy way. We proposed a percentile-based control method. For instance, when setting the MCTS participation proportion to 20%, we continuously collect η values during the inference process and dynamically calculate the η-threshold by computing the 20th percentile of the collected values. When the current η is larger than the computed η-threshold, we drive the policy LM to choose action; otherwise, MCTS is used. At the same time, the current η is added to the collected list, and the new η-threshold will be calculated. The MCTS ratios shown in Table 2 represent the actual MCTS participation rates obtained under settings of 0%, 25%, 50%, 75%, and 100% participation rates, respectively. Although this method lacks precise control, it is easy to implement and performs well in experiments.

Training Phase	Hyperparameter	Value
PT	Batch Size	8
	Training Epochs	5
	Learning Rate	6e-6
	Max Sequence Length	512
	Discount Factor	0.999
	Loss Weight λ_1	10.0
SPT	Training Epochs	5
	Learning Rate	1e-6
	Max Conversation Turn	8
	Discount Factor	0.999
	Loss Weight λ_2	1.0
	Training Size per Epoch	100

Table 5: Hyper-parameter settings in two-stage training phases for ESConv.

Training Phase	Hyperparameter	Value
PT	Batch Size	8
	Training Epochs	10
	Learning Rate	1e-5
	Max Sequence Length	512
	Discount Factor	0.999
	Loss Weight λ_1	10.0
SPT	Training Epochs	3
	Learning Rate	1e-5
	Max Conversation Turn	8
	Discount Factor	0.999
	Loss Weight λ_2	10.0
	Training Size per Epoch	100

Table 6: Hyper-parameter settings in two-stage training phases for CIMA.

C More Implementation Details

The training process for policy LM comprises two phases: offline RL-based pretraining (PT) and MCTS-guided self-play training (SPT). In the pretraining phase, DPDP is trained on the training set with collected rewards for each turn, and the checkpoint is saved based on the best performance observed on the validation set. Subsequently, during the MCTS-guided self-play training, cases from the training set are randomly sampled for online training. Due to cost considerations, we did not conduct an extensive hyperparameter search on ESConv.

Training Phase	Hyperparameter	Value
PT	Batch Size	8
	Training Epochs	10
	Learning Rate	6e-6
	Max Sequence Length	512
	Discount Factor	0.999
	Loss Weight λ_1	1.0
SPT	Training Epochs	3
	Learning Rate	1e-6
	Max Conversation Turn	5
	Discount Factor	0.999
	Loss Weight λ_2	1.0
	Training Size per Epoch	100

Table 7: Hyper-parameter settings in two-stage training phases for CraisglistBargain.

We mainly used the primary hyperparameters from PPDPP on ESConv and only conducted a search for learning rates within the range of [1e-6, 6e-6, 1e-5] on CIMA. The hyperparameters employed in our experiments are exhaustively detailed in Table 5, 6, and 7. All experiments are executed on a server equipped with 1 NVIDIA GeForce RTX 3090 GPU. Pretraining needs 128/6/210 minutes and self-play training needs 5/3/15.5 hours for ESConv/CIMA/CraisglistBargain.

D Details on MCTS Planner

We implement the MCTS planner referred to GDPZero (Yu et al., 2023). As supplementary, we introduce the procedures of MCTS here. For each node state in the MCTS tree, we store the sequence of actions that reached the current turn i as $s_i^{tr} = (a_0, ..., a_i)$. Given an initial conversation state of s_0, the MCTS planner searches for the next best action by iteratively performing action **selection**, search tree **expansion**, action **evaluation**, and **backpropagation** to update tree statistics. After n simulations, MCTS predicts the next best action for s_0. This process continues until the goal or the maximum number of dialogue turns is reached. Each of the four phases of MCTS is described below.

Selection Given a tree state s^{tr}, the action a^* with the highest Predictor Upper Confidence Tree Bound (PUCT) (Silver et al., 2017b) is selected to traverse the tree:

$$\begin{aligned}&\text{PUCT}(s^{tr}, a)\\ =&Q(s^{tr}, a) + c_p \cdot p(a|s^{tr}) \cdot \frac{\sqrt{\sum_a N(s^{tr}, a)}}{1 + N(s^{tr}, a)},\end{aligned}$$

where N records the number of times the pair (s^{tr}, a) has been visited, and c_p is a hyperparameter controlling exploration, setting as 1.0 for ESConv and CIMA. $p(a|s^{tr})$ is the prior action distribution.

In this study, we use policy LM to provide the prior probability for MCTS. The reasons lie in three points: (1) We hope to inject the domain knowledge learned by policy LM from the training set into MCTS;(2) We also hope to form an iterative optimization logic, that is, MCTS guides policy LM training in the self-play training, while a more capable policy LM can help MCTS find the next action; (3) We hope to reduce the number of calls to ChatGPT and thus reduce the cost.

Since future simulations require a specific dialogue history, we apply the previous simulation if the state is experienced before simulations, or generate a new simulation based on selected dialogue history h^{tr} by prompting. We repeat this process until s^{tr} becomes a leaf node.

Expansion Once a leaf node is reached, we utilize the policy LM to generate a prior policy, as previously described. Additionally, each node s^{tr} is initialized with $Q(s^{tr}, \cdot) = Q_0$, where Q_0 serves as a hyperparameter regulating exploration.

Evaluation The value of a state $v(s^{tr})$ is modeled based on the probability that its dialogue context can lead to task success. To assess dialogue states, we employ the method described in Section 3.2, wherein an LLM is prompted l times to estimate the user's state, and each comment is mapped to a pre-defined score.

Backpropagation At the conclusion of each search iteration, we systematically update the statistical attributes of every node traversed along the search path:

$$\begin{aligned}N(s^{tr}, a) &\leftarrow N(s^{tr}, a) + 1\\ Q(s^{tr}, a) &\leftarrow Q(s^{tr}, a) + \Delta Q(s^{tr}, a)\\ \Delta Q(s^{tr}, a) &= \frac{v(s^{tr} - Q(s^{tr}, a))}{N(s^{tr}, a)}.\end{aligned} \tag{7}$$

Prediction Upon the completion of all simulations, we designate the optimal action as $a^* = \arg\max_a N(s_0^{tr}, a)$, determined by the frequency of visitation for each action, where s_0^{tr} represents the root node of the tree.

E Human Evaluation Instruction

Regarding ESConv, we evaluate the responses from three primary perspectives and an overall perspective, as outlined below:

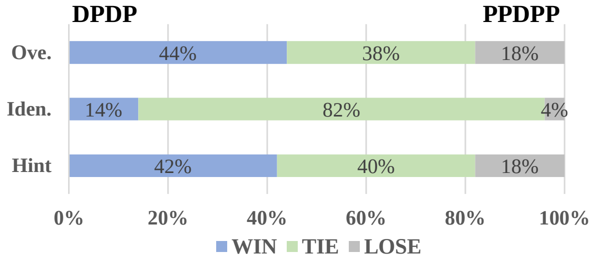


Figure 5: Human evaluation results on CIMA.

- Suggestion: Which assistant provides more helpful suggestions for solving the problem?
- Comforting: Which assistant is more skilled at comforting you?
- Identification: Which assistant is more helpful in exploring and identifying the problem?
- Overall: Which assistant can better solve the patient's problem?

As for CIMA, we also measure two main perspectives and an overall perspective of the responses as follows:

- Hint: Which assistant provides more helpful hints for translating correctly?
- Identification: Which assistant is better able to identify students' translation errors?
- Overall: Which assistant can better teach the student?

F More Human Evaluation

As a supplementary experiment, we also provide human evaluations conducted on CIMA, although evaluation biases on CIMA are not pronounced. The results are shown in Figure 5. The human evaluation results indicate that DPDP (System 1) outperforms, primarily evidenced by its more timely provision of prompts to students, as well as its increased guidance and flexibility in methodology.

G Prompting Details

In this section, we present the prompting details in our implementation. All the prompts we used are consistent with those in PPDPP.

G.1 Assistant Simulation

We will begin by delineating the specifics of the role-playing prompts utilized by the dialogue systems to generate assistant responses. This entails the utilization of dialogue strategy prompts, exemplified by [action], to direct the subsequent action within the dialogue.

ESConv In the emotional support dialogues, the assistant plays the role of a therapist, helping the patient mitigate emotional distress and address personal challenges. Each dialogue begins with the user expressing their concerns, described by [situation], which sets the specific background for the conversation. Detailed prompt is listed in Table 8

CIMA In the tutoring dialogues, the assistant acts as a teacher, guiding the student in translating English sentences into Italian. Each dialogue begins with a translation exercise, denoted by [exercise], and includes the student's specific difficulties with the exercise, represented by [situation], which provides a unique background for the conversation. Detailed prompt is presented in Table 9

CraisglistBargain In the negotiation dialogues, the assistant acts as the buyer, negotiating with the seller for a lower item price. Each scenario includes an item name [item name] and an item description [item description] to provide context for the negotiation. The buyer is given a target price to aim for, i.e., [buyer target price], and the negotiation starts at the listed item price, i.e., [seller target price]. Detailed prompt is presented in Table 10

G.2 User Simulation

Subsequently, we delineate the role-playing prompt designed to direct LLMs in simulating users, wherein the exclusion of dialogue strategy prompts ensures that simulated users respond solely to the dialogue history, abstaining from undertaking specific actions.

ESConv Within the domain of emotional support dialogues, the assistant adopts the role of a patient seeking assistance from the therapist. The prompt includes specifications of the emotion type [emotion type] and the problem type [problem type].

CIMA In the context of tutoring dialogues, the assistant assumes the role of a student tasked with acquiring the skill of translating English sentences into Italian. Since LLMs do well in translation, we further instruct them to forget the translation of the discussed exercise.

CraisglistBargain In the negotiation dialogues, the assistant takes on the role of the seller, bargaining with the buyer for a higher item price.

G.3 Reward Prompting

Concerning distinct conversational objectives, the prompts devised for the reward model are tailored to evaluate the extent of goal fulfillment.

System	Now enter the role-playing mode. In the following conversation, you will play as a therapist in a counseling conversation with a patient.
User	You are the therapist who is trying to help the patient reduce their emotional distress and help them understand and work through the challenges. Please reply with only one short and succinct sentence. [action] Are you ready to play the game?
Assistant	Yes, I'm ready to play the game!
User	[situation]

Table 8: Prompts for assistant simulation on ESConv.

System	Now enter the role-playing mode. In the following conversation, you will play as a teacher in a tutoring conversation with a student.
User	You are the teacher who is trying to teach the student to translate "[exercise]" into Italian. Please reply with only one short and succinct sentence. Please do not tell the student the answer or ask the student about other exercises. [action] Now ask me an exercise.
Assistant	Please translate "[exercise]" into Italian.
User	[situation]

Table 9: Prompts for assistant simulation on CIMA.

System	Now enter the role-playing mode. In the following conversation, you will play as a buyer in a price bargaining game.
User	You are the buyer who is trying to buy the [item name] with the price of [buyer target price]. Product description: [item description] Please reply with only one short and succinct sentence. [action] Now start the game.
Assistant	Hi, how much is the [item name]?
User	Hi, this is a good [item name] and its price is [seller target price].

Table 10: Prompts for assistant simulation on CraisglistBargain.

ESConv Given that the ultimate aim of emotional support dialogues is to address the patient's emotional issues comprehensively, we have structured four distinct levels of rewards to gauge the progression of these dialogues, as delineated in Table 14.

CIMA Given that the primary objective of tutoring dialogues is to instruct the student in correctly addressing the exercise, we have devised four tiers of rewards to evaluate the advancement of the tutoring dialogue, as outlined in Table 15.

CraisglistBargain Given that the goal of the negotiation dialogues is to reach a deal and maximize the benefit for the assistant, the reward model needs to first assess whether the user and the assistant have reached a deal, and then extract the final deal price to measure the benefit, as outlined in Table 16.

G.4 Strategy Prompting

Here, we present the mapping of dialogue strategies to their corresponding natural language prompts, utilized as [action] to direct the actions undertaken by the dialogue system. All prompts are consistent with PPDPP (Deng et al., 2023b).

ESConv ESConv is annotated with 8 emotional support strategies. In Table 17, we present these strategies alongside their corresponding natural language prompts tailored for LLMs.

CIMA The CIMA dataset is annotated with five tutoring strategies. In Table 19, we present these strategies along with their natural language prompts specifically crafted for LLMs.

CraisglistBargain The CraisglistBargain dataset is annotated with 11 negotiation strategies. In Table 19, we present these strategies along with their natural language prompts specifically crafted for

LLMs.

H Example Conversations

We present sample conversations generated by various dialogue systems interacting with the same user simulator under the same case in ESConv (Liu et al., 2021). We use the same case applied in the example demonstration of PPDPP (Deng et al., 2023b). Therefore, the examples for all baselines are from PPDPP. Finally, we provide conversations simulated using DPDP (policy LM) as the policy planner. Tables 20, 21, 22, 23, 24, and 25 display examples of emotional support conversations wherein the patient encounters a job crisis issue and experiences fear, necessitating resolution by the dialogue system.

System	Now enter the role-playing mode. In the following conversation, you will play as a patient in a counseling conversation with a therapist.
User	You are the patient who is looking for the help from the therapist, because you have the emotional issue about [emotion type] regarding [problem type]. Please reply with only one short and succinct sentence. Now tell me your issue.
Assistant	[situation]

Table 11: Prompts for user simulation on ESConv.

System	Now enter the role-playing mode. In the following conversation, you will play as a student who does not know Italian in a tutoring conversation with a teacher.
User	You are the student who is trying to translate an English sentence into Italian. You don't know the translation of "[exercise]" in Italian. Please reply with only one short and succinct sentence. Are you ready to play the game?
Assistant	Yes, I'm ready to play the game!
User	Please translate "[exercise]" into Italian.
Assistant	[situation]

Table 12: Prompts for user simulation on CIMA.

System	Now enter the role-playing mode. In the following conversation, you will play as a seller in a price bargaining game.
User	You are the seller who is trying to sell the [item name] with the price of [seller target price]. Product description: [item description] Please reply with only one short and succinct sentence. Are you ready to play the game?
Assistant	Yes, I'm ready to play the game!
User	Hi, how much is the [item name]?
Assistant	Hi, this is a good [item name] and its price is [seller target price].

Table 13: Prompts for user simulation on CraisglistBargain.

System	Given a conversation between a Therapist and a Patient, please assess whether the Patient' emotional issue has been solved after the conversation.
User	You can only reply with one of the following sentences: No, the Patient feels worse. No, the Patient feels the same. No, but the Patient feels better. Yes, the Patient's issue has been solved. The following is a conversation about [emotion type] regarding [problem type]: [conversation] Question: Has the Patient's issue been solved? Answer:

Table 14: Prompts for reward model on ESConv.

System	Given a conversation between a Teacher and a Student, please assess whether the Student correctly translate the English sentence into Italian in the conversation.
User	Please assess whether the Student correctly translated the whole sentence of "[exercise]" into Italian in the conversation. You can only reply with one of the following sentences: No, the Student made an incorrect translation. No, the Student did not try to translate. No, the Student only correctly translated a part of "[exercise]". Yes, the Student correctly translated the whole sentence of "[exercise]". The following is the conversation: [conversation] Question: Did the Student correctly translate the whole sentence of "exercise]" into Italian? Answer:

Table 15: Prompts for reward model on CIMA.

System	Given a conversation between a Buyer and a Seller, please decide whether the Buyer and the Seller have reached a deal at the end of the conversation.
User	Please decide whether the Buyer and the Seller have reached a deal at the end of the conver -sation. If they have reached a deal, please extract the deal price as [price]. You can only reply with one of the following sentences: They have reached a deal at [price]. They have not reached a deal. The following is the conversation: Buyer: Can we meet in the middle at $15? Seller: Sure, let's meet at $15 for this high-quality balloon. Question: Have they reached a deal? Answer: They have reached a deal at $15. The following is the conversation: Buyer: That's still a bit high, can you go any lower? Seller: Alright, I can sell it to you for $15. Question: Have they reached a deal? Answer: They have not reached a deal. The following is the conversation: [conversation] Question: Have they reached a deal? Answer:

Table 16: Prompts for reward model on CraisglistBargin.

Dialogue Strategy	Natural Language Form.
Question	Please ask the Patient to elaborate on the situation they just described.
Self-disclosure	Please provide a statement relating to the Patient about the situation they just described.
Affirmation and Reassurance	Please provide affirmation and reassurance to the Patient on the situation they just described.
Providing Suggestions	Please provide suggestion to the Patient on the situation they just described.
Reflection of feelings	Please acknowledge the Patient's feelings about the situation they described.
Information	Please provide factual information to help the Patient with their situation.
Restatement or Paraphrasing	Please acknowledge the Patient's feelings by paraphrasing their situation.
Others	Please chat with the Patient.

Table 17: Mapping of emotional support strategies to natural language prompting.

Dialogue Strategy	Natural Language Form.
Hint	Please provide knowledge to the Student via a hint.
Open-ended Question	Please ask a question to the Student to determine the Student's understanding or continue the conversation.
Correction	Please correct the mistake or address the misconception the Student has.
Confirmation	Please confirm the student's answer or understanding is correct.
Others	Please chat with the Student without any pedagogical strategy.

Table 18: Mapping of pedagogical strategies to natural language prompting.

Dialogue Strategy	Natural Language Form.
Greetings	Please say hello or chat randomly.
Ask a question	Please ask any question about product, year, price, usage, etc.
Answer a question	Please provide information about the product, year, usage, etc.
Propose the first price	Please initiate a price or a price range for the product.
Propose a counter price	Please propose a new price or a new price range.
Use comparatives	Please propose a vague price by using comparatives with existing price.
Confirm information	Please ask a question about the information to be confirmed.
Affirm confirmation	Please give an affirmative response to a confirm.
Deny confirmation	Please give a negative response to a confirm.
Agree with the proposal	Please agree with the proposed price.
Disagree with a proposal	Please disagree with the proposed price.

Table 19: Mapping of negotiation strategies to natural language prompting.

BACKGROUND INFORMATION
Emotion Type: Fear **Problem Type:** Job Crisis **Situation:** I think I will be losing my job soon. I just read an email talking about the need for us to cut costs and also how we have not got any support from the government.
STANDARD
Patient: I think I will be losing my job soon. I just read an email talking about the need for us to cut costs and also how we have not got any support from the government. Therapist: It sounds like you're feeling anxious and uncertain about your job security due to the current financial situation. Patient: Yes, I am feeling extremely anxious and uncertain about my job and financial stability. Therapist: I understand that this situation is causing you a lot of anxiety and uncertainty. Patient: Yes, it's been keeping me up at night and I can't stop worrying about what will happen if I lose my job. Therapist: It's understandable that the fear of losing your job is causing you significant distress and affecting your sleep. Patient: Yes, it's consuming my thoughts and I feel overwhelmed by the fear of not being able to support myself and my family. Therapist: The fear of not being able to provide for yourself and your family is a heavy burden to carry, and it's important to address these concerns. Patient: I know, but I just don't know what to do. I feel stuck and helpless in this situation. Therapist: Feeling stuck and helpless is a common response to uncertainty, but there are steps we can explore together to help you regain a sense of control and explore potential options. Patient: I'm willing to try anything at this point. I just want to find a way to alleviate this fear and find some stability in my life. Therapist: I'm glad to hear that you're open to exploring different strategies. Let's start by identifying your strengths and resources that can help you navigate this challenging situation. Patient: I'm not sure what strengths and resources I have, but I'm willing to explore and discover them. Therapist: That's a great mindset to have. Together, we can explore and uncover your strengths and resources that will support you in finding stability and managing your fear. Patient: Thank you, I appreciate your support. I'm ready to begin this journey of self-discovery and finding ways to overcome my fear of job crisis.
Turns of Conversation: 7

Table 20: Example conversations with Standard prompts in ESConv.

BACKGROUND INFORMATION

Same as Table 20

ASK-AN-EXPERT

Patient: I think I will be losing my job soon. I just read an email talking about the need for us to cut costs and also how we have not got any support from the government.
Therapist: I understand that receiving that email has caused you a lot of worry and uncertainty about your job security.
Patient: Yes, it has. I feel anxious and scared about what will happen to me if I lose my job.
Therapist: It's completely normal to feel anxious and scared in this situation. Remember that you have faced challenges before and have the strength to navigate through this one as well.
Patient: I know, but I can't help but constantly worry about the future and what it holds for me.
Therapist: It's understandable that you're worried about the future, but try to focus on the present moment and take things one step at a time.
Patient: I'll try, but it's hard not to let my fear of job loss consume my thoughts and affect my daily life.
Therapist: I hear you, and it's important to acknowledge those fears. Patient: Thank you for understanding. I just need some guidance on how to manage this fear and find some peace of mind.
Patient: That sounds helpful. How can I start practicing mindfulness?
Therapist: You can start by setting aside a few minutes each day to sit quietly, focus on your breath, and observe your thoughts and sensations without getting caught up in them.
Patient: I will give it a try. Thank you for suggesting mindfulness as a way to manage my fear and find some peace.
Therapist: You're welcome! Remember, it takes time and practice, so be patient with yourself. I'm here to support you along the way.
Patient: Thank you, I appreciate your support. It means a lot to me.

Turns of Conversation: 7

Table 21: Example conversations with Ask-an-Expert in ESConv.

BACKGROUND INFORMATION

Same as Table 20

PROCOT

Patient: I think I will be losing my job soon. I just read an email talking about the need for us to cut costs and also how we have not got any support from the government.
Therapist: It sounds like you're feeling anxious and uncertain about your job security due to the current financial situation.
Patient: Yes, I am feeling extremely anxious and uncertain about my job and financial stability.
Therapist: It's completely understandable to feel anxious and uncertain in this situation, but remember that you have the ability to adapt and find new opportunities.
Patient: I know, but the fear of not being able to support myself and my family is overwhelming.
Therapist: I can understand how overwhelming that fear can be, but it's important to remember that you have faced challenges before and have the strength to navigate through this one as.
Patient: I'm trying to remind myself of that, but the fear keeps consuming me and I can't seem to shake it off.
Therapist: It's normal for fear to consume us at times, but we can work together to develop strategies to manage and reduce that fear, allowing you to regain a sense.
Patient: I would really appreciate your help in developing those strategies to manage my fear and regain a sense of control.

Turns of Conversation: 4

Table 22: Example conversations with PROCOT in ESConv.

BACKGROUND INFORMATION
Same as Table 20
ICL-AIF
Patient: I think I will be losing my job soon. I just read an email talking about the need for us to cut costs and also how we have not got any support from the government. Therapist: It sounds like you're feeling worried and uncertain about your job security. Patient: Yes, I am extremely worried about what will happen if I lose my job. I have bills to pay and a family to support. Therapist: I can understand why you would feel so concerned about the potential impact losing your job could have on your financial stability and your ability to provide for your family. Patient: Yes, it's a constant source of anxiety for me. I feel like I'm constantly on edge, waiting for the other shoe to drop. Therapist: It's completely understandable that you would feel on edge and anxious given the uncertainty surrounding your job situation. It's important to acknowledge and address these feelings. Patient: I just don't know how to cope with this fear and uncertainty. It's consuming my thoughts and affecting my overall well-being. Therapist: I hear you, and it's important to find healthy coping strategies to manage your fear and uncertainty. I'm here to support you in developing those strategies. Patient: Thank you, I appreciate your support. I just need some guidance on how to navigate through this difficult time and find some peace of mind. Therapist: I'm here to help you navigate through this difficult time. Together, we can explore strategies to find peace of mind and manage the challenges you're facing. Patient: That means a lot to me. I'm ready to work on finding peace of mind and managing my fears about the job crisis.
Turns of Conversation: 5

Table 23: Example conversations with the ICL-AIF method in emotional support dialogues.

BACKGROUND INFORMATION
Same as Table 20
PPDPP
Patient: I think I will be losing my job soon. I just read an email talking about the need for us to cut costs and also how we have not got any support from the government. Therapist: I understand that receiving that email about potential job loss is causing you a lot of distress. Patient: Yes, it's been causing me a lot of anxiety and fear about my future. Therapist: It's completely normal to feel anxious and fearful in this situation, but remember that you have the strength and resilience to navigate through it. Patient: Thank you, I appreciate your reassurance. Therapist: You're welcome. Remember, you are not alone in this, and we will work together to find ways to cope and navigate through this challenging time. Patient: That's comforting to hear. I'm ready to work on finding ways to cope with my fear and uncertainty about the job crisis.
Turns of Conversation: 3

Table 24: Example conversations with the PPDPP method in emotional support dialogues.

BACKGROUND INFORMATION
Same as Table 20
DPDP (System 1)
Patient: I think I will be losing my job soon. I just read an email talking about the need for us to cut costs and also how we have not got any support from the government. Therapist: It sounds like you're feeling anxious and uncertain about your job security. Patient: Yes, I am feeling extremely anxious and uncertain about my job security. I can't stop worrying about what will happen if I lose my job. Therapist: It's understandable to feel anxious in this situation. Let's explore ways to manage your anxiety and develop a plan for different scenarios. Patient: Thank you, I really need help in managing my anxiety and finding ways to cope with the uncertainty. Therapist: I'm here to support you. Let's start by focusing on self-care strategies that can help reduce your anxiety and promote overall well-being. Patient: That sounds like a good place to start. What self-care strategies do you recommend for managing anxiety in the face of job uncertainty? Therapist: Some self-care strategies that may help manage anxiety during job uncertainty include maintaining a routine, practicing relaxation techniques such as deep breathing or meditation, engaging in physical exercise, seeking social support, and engaging in activities that bring you joy and relaxation. Patient: Thank you for those suggestions. I will try to incorporate those self-care strategies into my daily routine. Therapist: That's great to hear! Remember to be patient with yourself as you navigate through this challenging time. Is there anything else you would like to discuss or any specific concerns you would like to address? Patient: I would also like to discuss strategies for building resilience and exploring potential career options in case I do lose my job.
Turns of Conversation: 5

Table 25: Example conversations with the DPDP (System 1) method in emotional support dialogues.