A General Highly Accurate Online Planning Method Integrating Large Language Models into Nested Rollout Policy Adaptation for Dialogue Tasks

Hui Wang[1, 2], **Fafa Zhang**[1, 2], **Xiaoyu Zhang**[1, 2], **Chaoxu Mu**[1, 2, 3*]
[1]School of Artificial Intelligence, Anhui University
[2]Anhui Provincial Key Laboratory of Security Artificial Intelligence, Anhui University
[3]Pengcheng Laboratory, Shenzhen, China
{h.wang.13, zhangxiaoyu}@ahu.edu.cn, wa23301160@stu.ahu.edu.cn, cxmu@tju.edu.cn

Abstract

In goal-oriented dialogue tasks, the main challenge is to steer the interaction towards a given goal within a limited number of turns. Existing approaches either rely on elaborate prompt engineering, whose effectiveness is heavily dependent on human experience, or integrate policy networks and pre-trained policy models, which are usually difficult to adapt to new dialogue scenarios and costly to train. Therefore, in this paper, we present Nested Rollout Policy Adaptation for Goal-oriented Dialogue (NRPA-GD), a novel dialogue policy planning method that completely avoids specific model training by utilizing a Large Language Model (LLM) to simulate behaviors of user and system at the same time. Specifically, NRPA-GD constructs a complete evaluation mechanism for dialogue trajectories and employs an optimization framework of nested Monte Carlo simulation and policy self-adaptation to dynamically adjust policies during the dialogue process. The experimental results on four typical goal-oriented dialogue datasets show that NRPA-GD outperforms both existing prompt engineering and specifically pre-trained model-based methods. Impressively, NRPA-GD surpasses ChatGPT and pre-trained policy models with only a 0.6-billion-parameter LLM. The proposed approach further demonstrates the advantages and novelty of employing planning methods on LLMs to solve practical planning tasks.

Introduction

With the emergence of Large Language Models (LLMs) (Zhou et al. 2024) represented by ChatGPT (Synekop et al. 2024), in the field of Natural Language Processing (NLP) (Qin et al. 2024), there are many key technological breakthroughs for dialogue tasks (Algherairy and Ahmed 2024). Currently, LLM integration methods have significantly improved overall performance for goal-oriented dialogue tasks (Deng et al. 2025a). Goal-oriented dialogue tasks are designed to help users complete specific tasks with clear task objectives (Li 2024; Deng et al. 2024), such as bargaining, persuasion, negotiation, etc. (Deng et al. 2025b). Although LLM significantly improves the performance of dialogue systems, their actual performance is heavily dependent on well-designed prompts. Due to the prevalence of a large amount of contextually relevant content in dialogue scenarios, such as negotiation, persuasion, and emotional support, and other complex interaction contexts, these tasks often require the system to have the ability to dynamically adjust dialogue strategies. Therefore, it is particularly important to create a policy planner in goal-orientated dialogue tasks to analyze the contextual needs in real time according to the dialogue process, and maintain the consistency of the goal-orientated dialogue and the effectiveness of the policy throughout the interaction. This planning mechanism is designed not only to improve dialogue quality, but also to enable the system to better adapt to complex and changing dialogue situations.

Existing research has directly augmented LLM by designing heuristics or complex prompting processes, but such approaches are inefficient and have limited performance (Deng et al. 2023b). The subsequent emergence of policy planners combined with LLM, Yu et al. (Yu, Chen, and Yu 2023) use LLM to act as a prior policy, value function and complete tree search planning, resulting in improved performance, but still hardly satisfactory. To this end, He et al. (He et al. 2024) propose a two-stage framework: first, a policy model is trained to quickly give policies based on context; then a policy planner performs fine-grained planning for unfamiliar scenarios. This scheme achieves the best results in goal-oriented dialogues, but overall efficiency is still low, because a large amount of dedicated data is required for training the policy model and retraining is needed every time when the dialogue scenario is changed.

With the development of the field of artificial intelligence in decision optimization, Monte Carlo Tree Search (MCTS) combined with neural networks demonstrated the power of MCTS in decision optimization(Wang et al. 2016; Pavirani et al. 2024). In recent years, researchers have begun to explore the combination of MCTS and LLM, where MCTS provides structured search capabilities while LLM contributes generative capabilities. Zhang et al. (Zhang et al. 2024a) proposed a self-training method based on process reward enhancement that automatically optimizes inference steps. In addition, Zheng et al.(Zheng et al. 2025) successfully solved complex optimization problems using MCTS heuristic rule generation. DeLorenzo et al.(DeLorenzo et al. 2024) then combined the prospective search with the hardware design to generate high-performance hardware design

*Corresponding author: Chaoxu Mu.

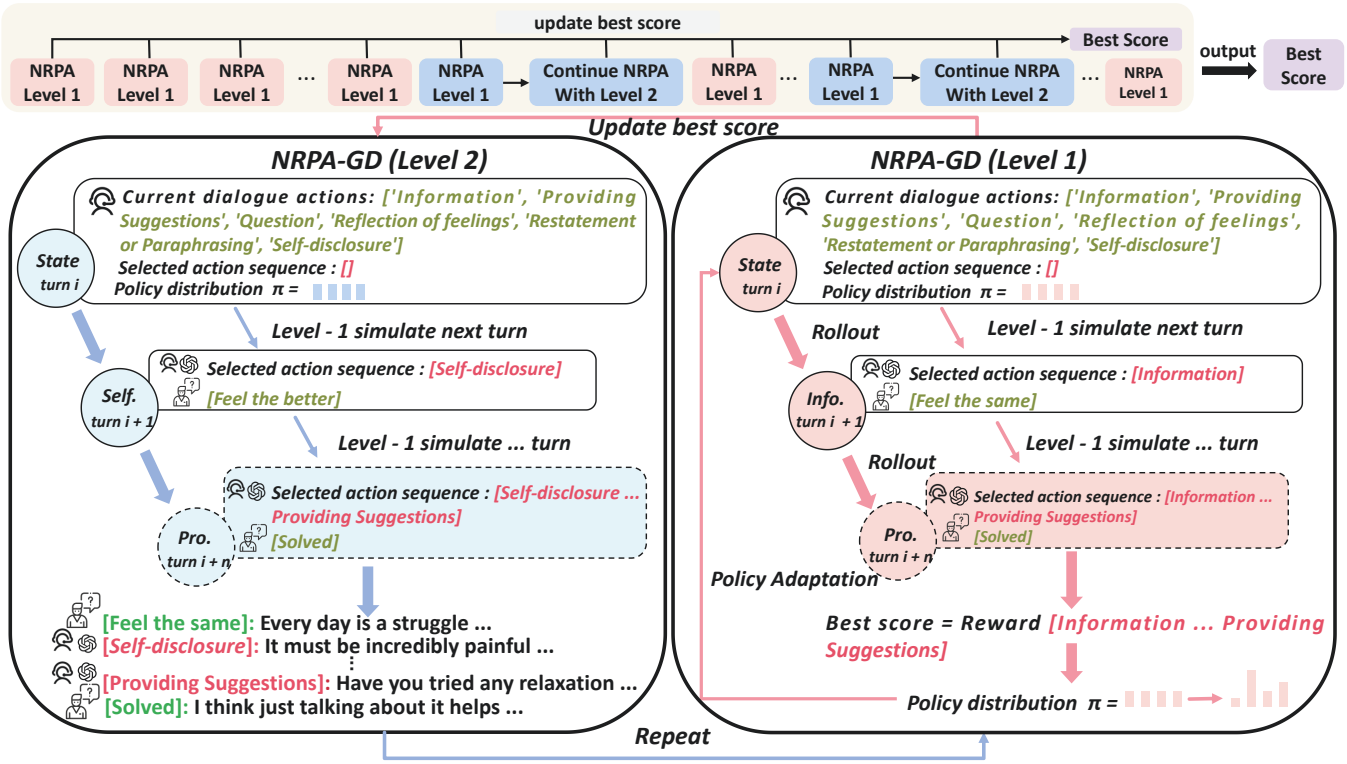


Figure 1: The NRPA-GD system uses a nested architecture to implement recursive policy optimization through two levels. In level 2, simulations are performed by recursively calling level 1 to optimize the policy distribution based on the reward values generated by policy choices, and the updated policies are passed to the next round of simulation iterations.

codes. Gan et al.(Gan et al. 2025) realized adaptive resource allocation for unsupervised tasks by dynamic collaboration of multiple intelligences. Inspired by MCTS as a dialog policy planner (Yu, Chen, and Yu 2023), we treat the goal-oriented dialog task as a combinatorial optimization problem (Wang, Zhang, and Mu 2025), which can be solved in a way similar to finding optimal paths under complex constraints. Many studies have shown that Nested Monte Carlo Search (NMCS) (Cazenave 2009) outperforms MCTS in a variety of optimization problems, and NMCS is particularly efficient in complex decision-making and optimization tasks (Cazenave 2009) because it finds the optimal solution by randomly sampling within a multilayer decision tree. The algorithm has been successfully applied to one-player games(Méhat and Cazenave 2010), two-player games (Cazenave et al. 2016), and RNA backfolding problems (Portela 2018). In addition, Nested Rollout Policy Adaptation (NRPA), an advanced variant of NMCS, has demonstrated improved performance in many domains, and NRPA further improves decision quality by optimizing the policy adaptation process (Rosin 2011). Compared to NMCS, NRPA is particularly advantageous in scenarios that require fine-grained policy tuning and fast policy updates. NRPA has already achieved significant results on tasks such as Minimum Congestion Shortest Path Routing (Banner and Orda 2007), Traveling Salesman Problem with Time Windows (Dumas et al. 1995), and Snake-in-the-Box (Abbott and Katchalski 1988), so the paradigm is also applicable to policy planning of dialogue scenarios.

Therefore, in this study, we propose a novel approach to plan the goal-oriented dialogue task based on NRPA. The NRPA is unique in that it employs a nested policy optimization mechanism, which achieves the global optimization of the dialogue policy through a multi-level policy evaluation and adjustment process. The method constructs complete dialogue trajectories in each dialogue simulation and iteratively updates the strategies based on these trajectories. This planning approach does not require pre-training of dialogue strategy models, but rather optimizes the strategy selection through online simulation and adaptive adjustment, thus effectively improving the goal attainment rate while maintaining dialogue fluency. Our comprehensive evaluation on multiple proactive dialogue tasks shows that NRPA-GD significantly outperforms existing policy planners on all metrics, and meets or surpasses that of pre-trained models for offline reinforcement learning. Our contributions can be summarized as follows.

- We propose an online planner powered by LLMs for goal-oriented dialogue tasks based on NRPA, eliminating the need for any pre-training of specific dialogue strategy models.
- We introduce NRPA-GD as an efficient search methods, and our experimental results demonstrate that NRPA-GD significantly outperforms the MCTS-based approach

while maintaining comparable or superior dialogue quality.

- Experiments demonstrate that NRPA-GD significantly outperforms existing policy planners in all metrics and achieves equivalent or even better performance than offline reinforcement learning training models.

Related Work

Dialog planning involves identifying appropriate strategies to guide the dialog before each dialog response. Existing prompt engineering approaches like Ask-an-Expert (AnE) (Zhang, Naradowsky, and Miyao 2023) prompt experts through predefined instructions to provide advice aligned with standard reasoning protocols in mental health, while the dialogue model learns from context and history which pieces of policy to adopt and which to discard. Proactive Chain-of-Thought (ProCoT) (Deng et al. 2023a) inserts an explicit goal-planning chain into the prompt so the model first infers its objective and then decides whether to probe further or refuse, revealing both the potential and the limitations of current large models in proactive conversation.

When combining pre-trained models as planning strategies, the Plug-and-Play Dialogue Policy (PPDPP) (Deng et al. 2023b) decomposes the dialogue policy into plug-and-play miniature models, trains them with supervised and reinforcement learning (RL) and significantly outperforms existing prompt or feedback-based approaches. Inspired by dual-process theory, Dual-Process Dialogue Planning (DPDP) (He et al. 2024) splits the LLM dialogue policy into a fast intuitive path and a slow analytical path. An offline-RL lightweight policy handles familiar contexts; when the situation becomes complex or unknown, an MCTS depth search is triggered to refine the policy in real time. Tailored Strategic Planning (TRIP) (Zhang et al. 2024b) integrates user-level traits into policy formulation, fusing user profiles and live feedback during planning through a population-based generalization framework that simulates multi-user distributions with evolutionary RL. User-Tailored Dialogue Policy Planning (UDP) (He et al. 2025b) uses a diffusion model to build user portraits on the fly, and predicts user feedback via a Brownian-bridge mechanism, and prioritizes hard samples with active learning, embedding individual traits into the policy loop to enhance system adaptability. Latent Dialogue Policy Planning (LDPP) (He et al. 2025a) employs a variational autoencoder (VAE) to mine latent policies from real conversations and then trains a hierarchical policy planner offline in that latent space, avoiding simulation bias and validating the feasibility and efficiency of self-discovered policies from real data. However, relying directly on a large language model as the policy or patching it with MCTS remains brittle. PPDPP, DPDP, TRIP, UDP, and LDPP all depend on offline-RL strategies for planning.

In addition, Yu et al. (Yu, Chen, and Yu 2023) propose Goal-oriented Dialogue Planning with Zero training (GDP-Zero), an MCTS-based policy planning method in which a single large language model simultaneously acts as a prior strategy, a value function, a user simulator, and a system model for unseen dialogue settings. Similarly, NRPA forgoes offline-RL entirely but refines policy from coarse to fine through an online multi-level recursive search (Rosin 2011) which achieve higher performance than MCTS in many domains. Therefore, we integrate LLMs into NRPA, to tackle the dialog tasks in this work.

Method

In this work, we propose a dialogue strategy planner based on the NRPA algorithm. This planner utilizes a zero-shot model training paradigm, performing multi-level exploration by prompting an LLM during decision-making. This process enables simulation of user-system responses, evaluation of current task progress, and optimization of the selection probability for the next dialogue act. As shown in Figure 1, within the NRPA-GD framework, the simulation process is set to a maximum level of 2. This process operates through recursive calls, optimizing the policy distribution based on the reward values obtained from policy selection. The updated policy is then passed to the next round of simulation iterations. Using the ESConv dataset as an example, the simulation terminates when the user action is "Solved" or the maximum dialogue turns are reached, at which point the best score is updated.

Problem Definition

To introduce the NRPA method for dialogue policy planning, we first formulate the planning task as a Markov Decision Process (MDP). A dialogue consisting of t turns between a user and a system can be represented as:

$$h_t = (u_0^{sys}, u_1^{usr}, u_1^{sys}, \ldots, u_t^{usr}, u_t^{sys}) \tag{1}$$

where u_i^{sys} is the system's response at turn i, and u_i^{usr} is the user's utterance at turn i. Each system response u_i^{sys} is associated with a specific dialogue act a_i^{sys}. Similar to (Yang, Li, and Guo 2021; Wang et al. 2020), we define the task of planning the next a^{sys} as an MDP problem $\langle S, A, T, R, \gamma \rangle$. The system's dialogue act a_i^{sys} at turn i represents an action $a_i \in A$. The corresponding dialogue history up to turn i,

$$s_i = (u_0^{sys}, u_1^{usr}, u_1^{sys}, \ldots, u_{i-1}^{usr}, u_{i-1}^{sys}) \tag{2}$$

represents a state $s_i \in S$. The transition function $T : S \times A \rightarrow S$ models the transition from state s_i to s_{i+1}, which occurs after the system executes action a_i to generate response u_i^{sys} and then receives the user's subsequent utterance u_{i+1}^{usr}. The reward function $R(s)$ evaluates the quality of the final state s of a simulated dialogue path. For instance, in an emotional support scenario, it measures whether the user's problem has been successfully resolved. We focus on evaluating the final state and implicitly incorporate discounting by applying a penalty to the dialogue length, thereby encouraging the model to achieve the goal in fewer turns. Specifically, if the dialogue is successfully solved, the reward value is 1, while a penalty of 0.001 times the dialogue turns is applied.

NRPA Planner

The algorithm 1 is used to simulate the dialogue in state s according to the current strategy π. The algorithm selects action a using softmax probabilistic sampling, calls the large

language model M_θ to generate system responses, and updates the state until the end of the dialogue to return the reward score and the sequence of actions.

The algorithm 2 implements adaptive updating of the strategy, receiving high-quality action sequences, and adjusting the weights of the strategy. For each action in the sequence, the mechanism of penalizing the global and rewarding the local is used to reduce the weights of all actions in proportion to the probability, and then significantly increase the weight of the current action, so that the strategy learns to optimize from the success experience.

The algorithm 3 implements multilevel strategy optimization, exploring dialogue paths, and selecting the best sequence update strategy through multiple playouts. The deeper levels of search provide more precise strategy guidance to the shallower levels, while the shallower levels are responsible for the broader exploration of the strategy space. This hierarchical optimization strategy enables NRPA-GD to efficiently discover high-quality dialogue strategies with limited computational resources, and achieves a novel combination of local fine-grained search and global strategy optimization.

The effectiveness of MCTS depends heavily on the rollout policy used in the simulation phase, and past approaches have either used static uniform stochastic policies or relied on manually tailored heuristics for specific domains, which severely limits the efficiency of the search. The NRPA upgrades to a dynamic policy optimization framework based on the NMCS. Instead of explicitly unfolding the entire search tree node by node, the algorithm adjusts the weight of the rollout policy in real time through gradient ascent within each nested hierarchy, which exponentially amplifies the probability of generating a high return path. The probability of generating high-return paths is exponentially increased. Specifically, let the state of the first t step be s_t, and let the set of legitimate actions be denoted as $\mathcal{A}(s_t)$. We parameterize the strategy as a vector : $\pi \in \mathbf{R}^{|\mathcal{A}|}$, where the component $\pi(a)$ corresponds directly to the weight of action a. Given the optimal sequence of actions $(a_1, a_2, \ldots, a_T)$ for a high score rollout, the following updates are performed for each step t:

Calculate the softmax normalization factor.

$$z = \sum_{a' \in \mathcal{A}} e^{\pi(a')} \tag{3}$$

Algorithm 1: NRPA Playout(s, π)

$sequence \leftarrow \emptyset$
$s \leftarrow s_i$
while s is not terminal **do**
  $z \leftarrow \sum_{a' \in A} e^{\pi(a')}$
  Draw a with probability $\frac{1}{z} e^{\pi(a)}$
  $u^{sys} \leftarrow M_\theta(s \circ a)$
  $s \leftarrow s \cup \{u^{sys}\}$
  append a to $sequence$
$score \leftarrow R(s)$
return $(score, sequence)$

Algorithm 2: NRPA Adapt(π, $sequence$, α, s)

$\pi' \leftarrow \pi$
$s \leftarrow s_i$
for a in $sequence$ **do**
  $z \leftarrow \sum_{a' \in \mathcal{A}} e^{\pi(a')}$
  for $a' \in \mathcal{A}$ **do**
    $\pi'(a') \leftarrow \pi'(a') - \alpha \cdot \frac{1}{z} e^{\pi(a')}$
  $\pi'(a) \leftarrow \pi'(a) + \alpha$
  $s \leftarrow play(s, a)$
return π'

Calculate the probability of each action.

$$P(a) = \frac{e^{\pi(a)}}{z} \tag{4}$$

Update the weights for all actions $a' \in \mathcal{A}$, and add an extra α to the optimal action a.

$$\begin{cases} \pi(a') \leftarrow \pi(a') - \alpha \cdot \frac{1}{z} e^{\pi(a')}, & \forall a' \in \mathcal{A} \\ \pi(a) \leftarrow \pi(a) + \alpha \end{cases} \tag{5}$$

The net increase in weight of the optimal action a is $\alpha - \alpha \cdot \frac{1}{z} e^{\pi(a)} = \alpha(1 - P(a))$, and the net decrease in weight of the remaining actions is $\alpha \cdot \frac{1}{z} e^{\pi(a')} = \alpha \cdot P(a')$, which transforms the original random simulation that was performed blindly into an adaptive sampling that continuously concentrates on the optimal direction.

Experiments

In this section, we introduce the experiment settings.

Datasets

We evaluated the proposed framework on four active dialogue datasets: ESConv (Liu et al. 2021) (Emotional Support Dialogue) containing 8 predefined actions; CIMA (Stasaski, Kao, and Hearst 2020) (Teaching and Learning Dialogues)

Algorithm 3: NRPA($level$, π, s)

Require: Generative LLM M_θ
Require: Level $level$, Policy π
Require: Number of iterations N
Require: Action space $\alpha \in \mathcal{A}$
if $level = 0$ **then**
  return PLAYOUT(s, π)
else
  $bestScore \leftarrow -\infty$
  $bestSequence \leftarrow \emptyset$
  for $iteration = 1$ to N **do**
    $(score, sequence) \leftarrow$ NRPA($level - 1$, π, s)
    if $score > bestScore$ **then**
      $bestScore \leftarrow score$
      $bestSequence \leftarrow sequence$
    $\pi \leftarrow$ ADAPT(π, $bestSequence$, α, s)
  return $(bestScore, bestSequence)$

containing 5 dialogue actions; and P4G (Wang et al. 2019) (Persuasion for Good). These three are all collaborative dialogue tasks (participants' goals are aligned), while CraigslistBargain (He et al. 2018) (Price Negotiation) serves as a non-collaborative task (buyer seeks lowest price/seller seeks highest price) involving 11 buyer bargaining actions.

Baselines

Our aim is to demonstrate the effectiveness of Monte Carlo methods in a goal-oriented dialogue framework through NRPA. To this end, we systematically compare three classes of baseline methods. Dialogue models based on generalized fine-tuning techniques are represented by DialoGPT(Zhang et al. 2019), a pre-trained dialogue generation model whose core functionality is to automatically generate natural, coherent and informative replies given a dialog context. Prompt-based engineering approaches like Standard Prompt(He et al. 2024) drive LLM generation of replies through base prompts; Proactive(Deng et al. 2023a) and ProCoT(Deng et al. 2023a) introduce explicit goal-planning chains in the prompts; Ask-an-Expert (Zhang, Naradowsky, and Miyao 2023) uses predefined prompts to model experts' standard reasoning strategies; ICL-AIF(Fu et al. 2023) generates textual feedback for zero-parameter updating contextual learning through model self-gaming; and GDPZero(Yu, Chen, and Yu 2023) innovates by allowing large language models to assume multiple roles in tree search. Offline RL-based approaches include the fine-tuned small model scheme of PPDPP(Deng et al. 2023b), and the dual processing mechanism of DPDP(He et al. 2024), which combines offline RL training and real-time MCTS search optimization.

Evaluation Metrics

In the evaluation, we use three key metrics: Average Turns (AT) and Success Rate (SR). AT measures the efficiency of goal completion by calculating the average number of dialogue turns required to reach a goal. SR measures the effectiveness of goal completion by counting the percentage of successful goal completion within a predefined maximum number of turns. And SL (Sale-to-List Ratio) evaluates the buyer's transaction outcome. The higher the SL, the more the buyer benefits from the deal; if the deal fails, the SL is recorded as 0. The formula is defined as: *SL% = (deal price - seller target price)/(buyer target price - seller target price) price)*. The same evaluation method of GDP-Zero was used in the P4G dataset, where we extracted the first 20 dialogues from P4G and generated a total of 154 rounds for evaluation (Yu, Chen, and Yu 2023). Using chatgpt as a judge then prompts the ChatGPT judge to select the more persuasive response. In addition, we found bias in the direct assessment using LLM. Therefore, we also performed a human evaluation as a comparison. Based on the evaluation dimensions in (He et al. 2024), we designed a multi-dimensional evaluation framework for different datasets. For the ESConv dataset, we compare the four dimensions of Suggestion (Sug.), Identification (Ide.), Comforting (Com.), and Overall (Ove.); for the CIMA dataset, we compare the four dimensions of Hint, Identification (Ide.), and Overall (Ove.); the P4G dataset focuses on Motivation (Motiv.), Persuasion (Pers.), Emotional (Emo.), and Overall (Ove.); and for the CraigslistBargain dataset it focuses on the Reasonableness (Rea.), Effectiveness (Eff.), Deal Success (Dea.) and Overall (Ove.). Among them, the evaluation dimensions of ESConv and CIMA directly adopt the criteria in (He et al. 2024). In the evaluation process, each annotator needs to judge which one of the different levels of NRPA-GD performs better for each dimension and give a conclusion of win, lose, or tie. To ensure the objectivity and consistency of the evaluation, we provided detailed evaluation guidelines and examples for the annotators, and conducted labeling consistency tests before the formal evaluation. In the end, we summarized and averaged the assessment results of the three annotators to obtain a more reliable and objective assessment conclusion.

Results and Analysis

This section presents the experimental results and corresponding analysis.

Static Evaluation

Table 1 compares the static GDP-Zero evaluation results with different nested levels of NRPA-GD on the P4G and ESConv datasets. The backbone model is gpt-4o-mini and the evaluator is gpt-3.5-turbo. GDP-Zero employs MCTS as the planner and adopts the best-reported setting from (Yu, Chen, and Yu 2023) with 50 simulations. NRPA-GD uniformly sets the number of single iterations at each level to 10 and introduces an early-stopping mechanism to curb the deepening of the layers with exponential time overhead caused by the deepening of the layers. The experimental results show that when the nesting level of NRPA-GD is 1, its planning response on P4G is preferred in up to 68. 61%, with much less time consumption compared to GDP-Zero. And in ESConv, the win rate for level 1 is up to 63. 40%, with the lowest time cost. By further increasing the level to 2, the computational cost increases significantly, but the best performance achieved on both datasets.

Automatic Evaluation

As shown in Table 2, the overall performance of NRPA-GD is accompanied by a tier improvement. In CIMA, the SR of both Level 1 and Level 2 reaches 100%, while the AT decreases from the best 2.24 to 1.03, with a decrease of more than 54%, which is better than the previous pre-trained policy models or the MCTS baseline. This suggests that a multilayer search can quickly target the optimal policy in tasks with small action space. In ESConv, although the AT (3.65) of Level-2 is still higher than the DPDP (2.13), the SR reaches 100%, which is critical in emotional support scenarios where the problem is solved. In CraigslistBargain, Level-1 increased SL from 0.4108 to 0.6371 while maintaining SR at 100% and AT at 3.11, and Level 2 decreased SL to 0.5161, but the AT was further decreased to 2.61, indicating that after increasing the number of simulations, NRPA-GD achieved a balance between AT, SR and SL. NRPA-GD achieves a success rate of 100% in all three datasets by expanding the

Method	P4G		ESConv	
	Run Time	Win Rate vs. ChatGPT	Run Time	Win Rate vs. ChatGPT
GDP-Zero (Yu, Chen, and Yu 2023)	636 s	$58.66\% \pm 2.73\%$	431 s	$52.00\% \pm 0.52\%$
NRPA-GD (Level 1)	**239 s**	$68.61\% \pm 1.84\%$	**287 s**	$63.40\% \pm 0.52\%$
NRPA-GD (Level 2)	1039 s	**77.49% ± 2.25%**	1033 s	**74.50% ± 2.26%**

Table 1: Static evaluation with ChatGPT as backbone and judge. Results are given as $\mu \pm \sigma$ repeated over three runs.

Models	ESConv		CIMA		CraigslistBargain		
	AT↓	SR↑	AT↓	SR↑	AT↓	SR↑	SL↑
DialoGPT (Zhang et al. 2019)	5.31	0.7538	5.43	0.4956	6.73	0.3245	0.2012
Standard (He et al. 2024)	5.10	0.7692	3.89	0.6903	6.47	0.3830	0.1588
AnE (Zhang, Naradowsky, and Miyao 2023)	4.76	0.8000	3.86	0.6549	5.91	0.4521	0.2608
Proactive (Deng et al. 2023a)	5.08	0.7538	4.84	0.5310	5.80	0.5638	0.2489
ProCoT (Deng et al. 2023a)	4.75	0.7923	4.58	0.5487	6.22	0.5319	0.2486
ICL-AIF (Fu et al. 2023)	4.69	0.8079	4.19	0.6106	6.53	0.3617	0.1881
PPDPP (Deng et al. 2023b)	4.56	0.8462	3.03	0.8407	5.62	0.6117	0.3376
DPDP (System 1) (He et al. 2024)	3.61	0.9000	2.24	0.9469	5.03	0.7447	0.4108
-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	2.78	0.9734	0.2728
DPDP (System 1&2)	**2.13**	0.9923	2.28	0.9823	-	-	-
NRPA-GD (Level 1)	3.85	**1.0000**	1.08	**1.0000**	3.11	**1.0000**	**0.6371**
NRPA-GD (Level 2)	3.65	**1.0000**	**1.03**	**1.0000**	**2.61**	**1.0000**	0.5161

Table 2: Automatic Evaluation Results with ChatGPT on ESConv, CIMA, and CraigslistBargain.

depth of search, and the focus on high-return dialogues can improve the surplus without sacrificing the deal completion rate.

Performance on Different LLMs

To further validate the effectiveness of the proposed framework under different model sizes, we use three different sizes of LLMs: gpt-4o-mini, llama-3.1-8b-chat, and qwen-3-0.6b. We evaluated the performance of NRPA-GD Level 1 and Level 2 respectively, and the results are shown in Table 3. With gpt- 4o-mini and llama-3.1-8b-chat, NRPA-GD shows a consistent performance improvement trend when the nesting level is increased, with both AT and SR improving and maintaining high returns on the CraigslistBargain dataset. This validates the stability and effectiveness of the NRPA-GD framework on medium- and large-scale models. For the qwen-3-0.6b model with the smallest parameter size, it is unable to complete the ESConv experimental task within the preset limit of dialogue rounds due to its parameter size limitation. However, the model still demonstrated the advantages of NRPA-GD on other datasets. In the CIMA data set, for both the AT and SR metrics of Level 2, the method achieves the desired high success rate performance with short rounds. And in the CraigslistBargain data set, the SL value of Level 2 also improves significantly to **0.6307**, which is comparable to the performance of gpt-4o-mini and gpt-3.5-turbo. It is evidenced that NRPA-GD is able to achieve the higher performance level than that of large parameter pre-trained models on specific tasks even with small parameter models.

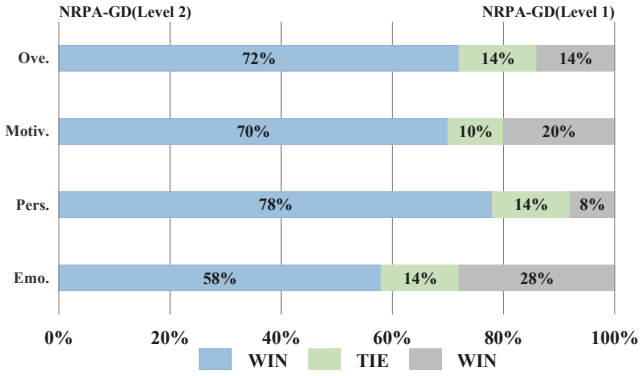
Human Evaluation

Based on previous research (Liu et al. 2021), LLM may lack rigorous criteria for persuasion components due to its possible preference for self-generated dialogues. In order to objectively assess the model performance, we randomly selected 50 sets of dialogues from each of the P4G and CIMA datasets for multidimensional human evaluation. During the evaluation process, we designed the corresponding evaluation dimensions for the specific task goals of different datasets. With an increase in search depth, the NRPA-GD in Figure 3 shows a significant performance improvement on both datasets. In the P4G task, it can provide emotional support and promote donation behavior more effectively, and in the CIMA task, the model can provide students with more accurate learning hints. More importantly, at the methodology level, the NRPA-GD model demonstrates stronger guidance and flexibility, and can maintain consistent and high-quality performance in different dialogue scenarios.

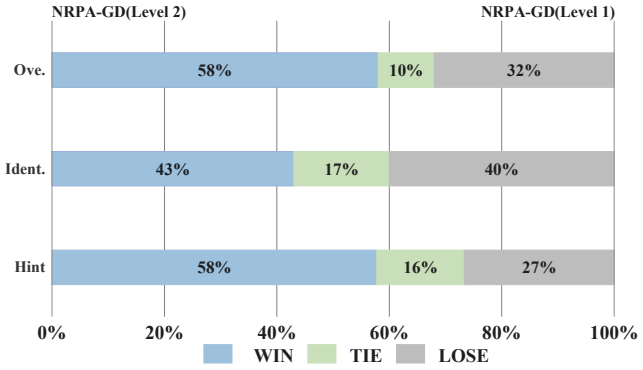
Similarly, we manually evaluated 50 dialogues in each of the ESConv and CraigslistBargain datasets. The evaluation results in Figures 3a and 3b reveal the performance of

LLM	Level	ESConv		CIMA		CraigslistBargain		
		AT↓	SR↑	AT↓	SR↑	AT↓	SR↑	SL↑
GPT-4o-mini	1	5.28	0.9461	2.04	**1.0000**	3.89	0.9894	0.6326
	2	4.17	**1.0000**	1.76	**1.0000**	**2.72**	**1.0000**	**0.6422**
Llama3.1-8b	1	3.98	**1.0000**	1.92	**1.0000**	3.54	**1.0000**	0.5166
	2	**3.65**	**1.0000**	1.69	**1.0000**	3.16	**1.0000**	0.4938
Qwen3-0.6b	1	-	-	1.22	**1.0000**	3.28	0.9894	0.6159
	2	-	-	**1.08**	**1.0000**	3.17	**1.0000**	0.6307

Table 3: Experimental results comparing LLM performance on ESConv, CIMA, and CraigslistBargain.



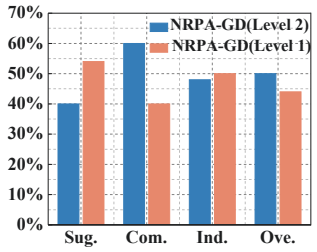
(a) P4G.



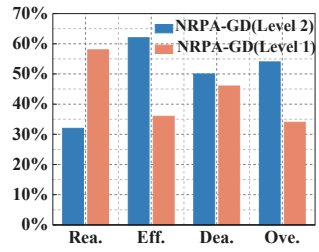
(b) CIMA.

Figure 2: Human evaluation results on P4G and CIMA.

the model at different search depths. In the ESConv task, NRPA-GD (Level 1) tends to provide concrete and feasible suggestions, showing a strong problem solving orientation, while NRPA-GD (Level 2) focuses more on building emotional empathy with patients, and the two levels are comparable in terms of their ability to understand the emotional state of patients and identify core problems. In CraigslistBargain, the models exhibited different strategy preferences. nRPA-GD (Level 1) demonstrated more rational negotiation behavior, while NRPA-GD (Level 2) was superior in practical results, being more effective in reaching satisfactory negotiation outcomes and facilitating successful transactions, reflecting stronger goal orientation and execution.



(a) ESConv.



(b) CraigslistBargain.

Figure 3: Human evaluation results on ESConv and CraigslistBargain.

Based on superior performance, NRPA-GD (Level 2) outperforms Level 1 in the overall evaluation, validating the effectiveness of the deep search strategy in complex dialogue tasks.

Conclusion

We propose NRPA-GD, an LLM-based online policy planning algorithm designed for goal-oriented dialogue systems. Moreover, it enables policy optimization without additional training. NRPA-GD's nested simulation mechanism enables dynamic exploration of multiple possible interaction paths during dialogue, while the adaptive update mechanism ensures that the policies can be optimized and adjusted based on real-time feedback. The experimental results show that the policies generated by NRPA-GD surpass the previous best system (DPDP), achieving stable and significant improvements in the evaluation metrics. NRPA-GD also surpasses ChatGPT even with an LLM with only 0.6B parameters, dramatically improving the intelligence of the dialogue system without increasing model parameters and training costs. For future work, it is necessary to explore more pruning methods to further improve the time efficiency of the proposed approach.

Acknowledgments

The authors acknowledge the financial support from National Natural Science Foundation of China, No. 62236002, and Hefei Key Science and Technology Special Projects under Grant 2024SZD006.

References

Abbott, H. L.; and Katchalski, M. 1988. On the snake in the box problem. *Journal of Combinatorial Theory, Series B*, 45(1): 13–24.

Alghera iry, A.; and Ahmed, M. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12): 6325–6351.

Banner, R.; and Orda, A. 2007. Multipath routing algorithms for congestion minimization. *IEEE/ACM Transactions on networking*, 15(2): 413–424.

Cazenave, T. 2009. Nested Monte-Carlo Search. In *IJCAI*, volume 9, 456–461.

Cazenave, T.; Saffidine, A.; Schofield, M.; and Thielscher, M. 2016. Nested Monte Carlo search for two-player games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

DeLorenzo, M.; Chowdhury, A. B.; Gohil, V.; Thakur, S.; Karri, R.; Garg, S.; and Rajendran, J. 2024. Make every move count: Llm-based high-quality rtl code generation using mcts. *arXiv preprint arXiv:2402.03289*.

Deng, Y.; Liao, L.; Chen, L.; Wang, H.; Lei, W.; and Chua, T.-S. 2023a. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

Deng, Y.; Liao, L.; Lei, W.; Yang, G. H.; Lam, W.; and Chua, T.-S. 2025a. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3): 1–45.

Deng, Y.; Liao, L.; Lei, W.; Yang, G. H.; Lam, W.; and Chua, T.-S. 2025b. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3): 1–45.

Deng, Y.; Liao, L.; Zheng, Z.; Yang, G. H.; and Chua, T.-S. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 807–818.

Deng, Y.; Zhang, W.; Lam, W.; Ng, S.-K.; and Chua, T.-S. 2023b. Plug-and-play policy planner for large language model powered dialogue agents. *arXiv preprint arXiv:2311.00262*.

Dumas, Y.; Desrosiers, J.; Gelinas, E.; and Solomon, M. M. 1995. An optimal algorithm for the traveling salesman problem with time windows. *Operations research*, 43(2): 367–371.

Fu, Y.; Peng, H.; Khot, T.; and Lapata, M. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Gan, B.; Zhao, Y.; Zhang, T.; Huang, J.; Li, Y.; Teo, S. X.; Zhang, C.; and Shi, W. 2025. MASTER: A Multi-Agent System with LLM Specialized MCTS. *arXiv preprint arXiv:2501.14304*.

He, H.; Chen, D.; Balakrishnan, A.; and Liang, P. 2018. Decoupling strategy and generation in negotiation dialogues. *arXiv preprint arXiv:1808.09637*.

He, T.; Liao, L.; Cao, Y.; Liu, Y.; Liu, M.; Chen, Z.; and Qin, B. 2024. Planning like human: A dual-process framework for dialogue planning. *arXiv preprint arXiv:2406.05374*.

He, T.; Liao, L.; Cao, Y.; Liu, Y.; Sun, Y.; Chen, Z.; Liu, M.; and Qin, B. 2025a. Simulation-free hierarchical latent policy planning for proactive dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24032–24040.

He, T.; Liao, L.; Liu, M.; and Qin, B. 2025b. Simulating before planning: Constructing intrinsic user world model for user-tailored dialogue policy planning. *arXiv preprint arXiv:2504.13643*.

Li, X. 2024. A review of prominent paradigms for llm-based agents: Tool use (including rag), planning, and feedback learning. *arXiv preprint arXiv:2406.05804*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Méhat, J.; and Cazenave, T. 2010. Combining UCT and nested Monte Carlo search for single-player general game playing. *IEEE Transactions on Computational Intelligence and AI in Games*, 2(4): 271–277.

Pavirani, F.; Gokhale, G.; Claessens, B.; and Develder, C. 2024. Demand response for residential building heating: Effective Monte Carlo Tree Search control based on physics-informed neural networks. *Energy and Buildings*, 311: 114161.

Portela, F. 2018. An unexpectedly effective Monte Carlo technique for the RNA inverse folding problem. *BioRxiv*, 345587.

Qin, L.; Chen, Q.; Feng, X.; Wu, Y.; Zhang, Y.; Li, Y.; Li, M.; Che, W.; and Yu, P. S. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Rosin, C. D. 2011. Nested rollout policy adaptation for Monte Carlo tree search. In *Ijcai*, volume 2011, 649–654.

Stasaski, K.; Kao, K.; and Hearst, M. A. 2020. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 52–64.

Synekop, O.; Lytovchenko, I.; Lavrysh, Y.; and Lukianenko, V. 2024. Use of Chat GPT in English for engineering classes: Are students' and teachers' views on its opportunities and challenges similar? *International Journal of Interactive Mobile Technologies*, 18(3).

Wang, F.-Y.; Zhang, J. J.; Zheng, X.; Wang, X.; Yuan, Y.; Dai, X.; Zhang, J.; and Yang, L. 2016. Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica*, 3(2): 113–120.

Wang, H.; Zhang, X.; and Mu, C. 2025. Planning of Heuristics: Strategic Planning on Large Language Models with Monte Carlo Tree Search for Automating Heuristic Optimization. *arXiv preprint arXiv:2502.11422*.

Wang, S.; Zhou, K.; Lai, K.; and Shen, J. 2020. Task-completion dialogue policy learning via Monte Carlo tree search with dueling network. In *Proceedings of the 2020*

conference on empirical methods in natural language processing (EMNLP), 3461–3471.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Yang, J.; Li, S.; and Guo, J. 2021. Multi-turn target-guided topic prediction with Monte Carlo tree search. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, 324–334.

Yu, X.; Chen, M.; and Yu, Z. 2023. Prompt-based Monte-Carlo tree search for goal-oriented dialogue policy planning. *arXiv preprint arXiv:2305.13660*.

Zhang, D.; Zhoubian, S.; Hu, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024a. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37: 64735–64772.

Zhang, Q.; Naradowsky, J.; and Miyao, Y. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. *arXiv preprint arXiv:2305.17878*.

Zhang, T.; Huang, C.; Deng, Y.; Liang, H.; Liu, J.; Wen, Z.; Lei, W.; and Chua, T.-S. 2024b. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. *arXiv preprint arXiv:2403.06769*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zheng, Z.; Xie, Z.; Wang, Z.; and Hooi, B. 2025. Monte carlo tree search for comprehensive exploration in llm-based automatic heuristic design. *arXiv preprint arXiv:2501.08603*.

Zhou, H.; Hu, C.; Yuan, Y.; Cui, Y.; Jin, Y.; Chen, C.; Wu, H.; Yuan, D.; Jiang, L.; Wu, D.; et al. 2024. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *IEEE Communications Surveys & Tutorials*.