E-NRPA: EMOTION-AWARE NESTED ROLLOUT POLICY ADAPTATION ON LARGE LANGUAGE MODELS FOR GOAL-ORIENTED DIALOGUE PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

User emotion influences strategy selection in goal-oriented dialogue, but existing online planning methods typically do not explicitly model changes in emotion. This paper proposes Emotion-Aware Nested Rollout Policy Adaptation for Goal-Oriented Dialogue Planning (E-NRPA), which incorporates current emotion and historical emotion trajectories identified by large language models into Monte Carlo simulation and policy update processes, enabling the planner to adjust dialogue strategies based on changes in user state. We conduct experiments on ESConv, ExTES, CraigslistBargain, and P4G. Results show that E-NRPA achieves stable performance on automatic metrics, with advantages in task success rate and negotiation outcome quality. Human evaluations indicate that E-NRPA is preferred in negotiation tasks and receives better ratings for conciseness, goal progress, and overall efficiency in emotional support dialogues. The incorporation of emotion trajectories further enhances the planner's sensitivity to changes in user state and improves its ability to adapt strategy selection dynamically. Compared with methods that rely only on the current dialogue state, E-NRPA can better capture how emotional evolution influences user goals and interaction intentions. These findings suggest that emotion-aware planning facilitates effective, goal-oriented dialogue strategy selection across different tasks.

1 INTRODUCTION

The rise of large language models (LLMs), exemplified by ChatGPT Zhao et al. (2023), has driven significant progress in natural language processing (NLP) and dialogue systems Algherairy & Ahmed (2024). Integrating LLMs has notably boosted the overall performance of goal-oriented dialogue tasks Deng et al. (2025), which aim to help users achieve specific objectives—such as price negotiation, persuasion, bargaining, and emotional support—within limited turns Deng et al. (2024); Liu et al. (2021); He et al. (2018). Despite these advances in fluency and task completion, the effectiveness of LLMs still heavily depends on meticulous prompt engineering. In practice, goal-oriented dialogue is embedded with rich contextual signals, where evolving user emotions critically shape interaction outcomes Wan et al. (2025). For example, in emotional support, a user's shift from frustration to hope may affect whether the core issue is resolved; in persuasion, user characteristics and emotional appeals can influence strategy effectiveness and dialogue outcomes Wang et al. (2019). Therefore, an effective dialogue strategy must not only guide the conversation toward its goal, but also perceive, track, and adapt to the user's emotional trajectory in real time.

Existing methods fall into two main categories, both of which have limitations in handling emotional dynamics. The first category includes prompt-engineering techniques such as Ask-an-Expert (AnE) Zhang et al. (2023) and Proactive Chain-of-Thought (ProCoT) Deng et al. (2023a). These methods rely on static heuristic instructions or explicit planning chains. While they improve strategic reasoning, they lack continuous emotional monitoring and cannot adapt to rapid emotional shifts during interaction. The second category consists of hybrid frameworks that combine pretrained policy models with planners, including Plug-and-Play Dialogue Policy Planner (PPDPP) Deng et al. (2023b) and Dual-Process Dialogue Planning (DPDP) He et al. (2024). PPDPP trains a lightweight plug-in policy planner through supervised fine-tuning and reinforcement learning, whereas DPDP combines an offline-trained policy model with an MCTS planner for complex or unfamiliar scenarios. More-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

over, they typically process emotional information implicitly through surface-level linguistic cues without explicitly maintaining the user's historical emotional states, which limits their ability to perform long-term policy optimization.

In recent years, decision-time planning has shown potential in combining Monte Carlo methods and their variants with LLMs Zhang et al. (2024a); Zheng et al. (2025); Yu et al. (2023). Decision-time planning refers to performing planning during inference, where the system evaluates possible future actions before selecting each response, rather than relying solely on a fixed policy learned offline. Goal-oriented Dialog Planning with Zero training (GDP-Zero) Yu et al. (2023) uses a single LLM to simultaneously serve as the prior policy, value function, user simulator, and system model. Nested Rollout Policy Adaptation (NRPA) Rosin (2011) is a simulation-based Monte Carlo planning method that uses nested rollouts and policy adaptation to iteratively improve action selection. Building on NRPA, Nested Rollout Policy Adaptation for Goal-oriented Dialogue (NRPA-GD) Wang et al. (2026) replaces the MCTS-based planning process with NRPA to achieve online policy adaptation with improved sample efficiency, without requiring offline reinforcement learning training. However, neither GDP-Zero nor NRPA-GD incorporates user emotion as a core component of the planning state. This omission limits the system's ability to generate responses that are both goal-oriented and emotionally appropriate, making it difficult to meet the requirements of a human-like, empathetic, and efficient dialogue system.

To address this gap, we propose Emotion-Aware Nested Rollout Policy Adaptation for Goal-Oriented Dialogue Planning, a novel online planning framework that seamlessly integrates LLMs with emotion recognition and historical emotion trajectory tracking into the NRPA architecture. Specifically, following each user response, the LLM infers the current emotional state and injects the real-time emotional label, along with the complete dialogue history and historical emotional trajectories, into the rollout simulation process, thereby indirectly influencing policy adaptation through the selected high-scoring trajectories. By incorporating emotional trajectories into Monte Carlo simulations and strategy adaptation processes, E-NRPA is able to adjust its dialogue strategies in response to changes in the user's state, while retaining the key feature of NRPA-based methods that they do not require offline training. Our contributions include: Our contributions are:

- We propose E-NRPA, an emotion-aware online dialogue planner that requires no offline reinforcement learning.
- We explicitly model current and historical user emotions to support task-oriented, emotionally aligned responses.
- We incorporate emotion trajectories into rollout simulation, enabling dynamic strategy adaptation to changes in user state.

2 Related Work

Existing dialogue policy planning methods can be roughly categorized into three categories. The first category focuses on user simulation and user modeling. Tailored Strategy Planning (TRIP) Zhang et al. (2024b) constructs multiple user simulators based on user profiles and resisting strategies, and uses large language models to infer users' mental states and possible future actions from historical dialogues. However, it relies on user simulation and reinforcement learning-based optimization, which may increase training cost and introduce a gap between simulated and real users. User-Tailored Dialogue Policy Planning (UDP) He et al. (2025b) uses diffusion models to generate dynamic user profiles and Brownian Bridges to predict user feedback, but its performance depends on the quality of profile generation and feedback prediction, with additional modeling complexity. NRPA-GD Wang et al. (2026) utilizes large language models to simulate both users and the system, and continuously updates policies through nested policy optimization, but repeated LLM-based simulations may lead to high interaction cost.

The second category focuses on action candidate generation and policy selection. DialogXpert Rakib et al. (2026) uses a frozen LLM to generate high-quality action candidates and selects the final action with a trained Q-network embedded in a BERT encoder, but it requires extra Q-network training and value estimation. PPDPP Deng et al. (2023b) adopts a plug-and-play policy optimization framework, using small pluggable language models trained via reinforcement learning and supervised learning as policy planners, but still requires training additional planner modules.

Algorithm 1 E-NRPA

Require: LLM M_θ, initial policy π, number of iterations N, learning rate α, action space $\mathcal{A}$, initial emotion-aware state $s = (H, E)$

function E-NRPA($level$, π, s)
  if $level = 0$ **then**
    return PLAYOUT(s, π)
  else
    $bestScore \leftarrow -\infty$
    $bestSequence \leftarrow \emptyset$
    for $iteration = 1$ to N **do**
      $(score, sequence) \leftarrow$ E-NRPA($level - 1$, π, s)
      if $score > bestScore$ **then**
        $bestScore \leftarrow score$
        $bestSequence \leftarrow sequence$
      $\pi \leftarrow$ ADAPT(π, $bestSequence$, α, s)
    return $(bestScore, bestSequence)$
function ADAPT(π, $sequence$, α, s)
  $\pi' \leftarrow \pi$
  $currentState \leftarrow s$
  for each action a in $sequence$ **do**
    $z \leftarrow \sum_{a' \in \mathcal{A}} e^{\pi'(a')}$
    for each action $a' \in \mathcal{A}$ **do**
      $\pi'(a') \leftarrow \pi'(a') - \alpha \cdot \frac{1}{z} e^{\pi'(a')}$
    $\pi'(a) \leftarrow \pi'(a) + \alpha$
    $currentState \leftarrow play(currentState, a)$
  return π'

The third category introduces reasoning, search, or latent policy modeling. DPDP He et al. (2024) combines a fast neural policy language model with a MCTS planner for complex scenarios, but MCTS-based planning can be computationally expensive. Latent Dialogue Policy Planning He et al. (2025a) mines latent policies from real-world dialogue logs and trains a high-level policy planner via Vector Quantized Variational Autoencoder-based latent policy discovery and offline hierarchical reinforcement learning, but its training pipeline is complex and data-dependent.

Existing methods improve dialogue policy planning from the perspectives of user simulation, action selection, search-based reasoning, and latent policy modeling, but they often rely on additional training, user simulation, or complex planning processes. Moreover, user emotional changes are usually treated as auxiliary information rather than being used to guide policy selection. To address this issue, our method explicitly incorporates users' historical emotions, current dialogue context, and emotional change trends into policy planning, aiming to improve emotion understanding, personalized response generation, and policy adaptation.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Building on existing work, we formulate the dialogue planning process as a Markov decision process (MDP), denoted as $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$, where $\mathcal{S}$ denotes the dialogue state space, $\mathcal{A}$ denotes the dialogue action space, $\mathcal{T}$ denotes the state transition function, and $\mathcal{R}$ denotes the reward function. At turn t, the current state $s_t \in \mathcal{S}$ contains the dialogue history and contextual information accumulated up to that turn. The output space of the planner consists of a set of predefined dialogue strategies derived from existing research, each accompanied by a pre-designed natural language instruction. Given the current state s_t, the planner selects an action $a_t \in \mathcal{A}$, after which the transition function $\mathcal{T}$ produces the next state s_{t+1} and the system receives an immediate reward r_t. Following NRPA-GD, the reward function $\mathcal{R}$ is calculated based on the final dialogue state, the total number of dialogue turns, and the corresponding penalty mechanism.

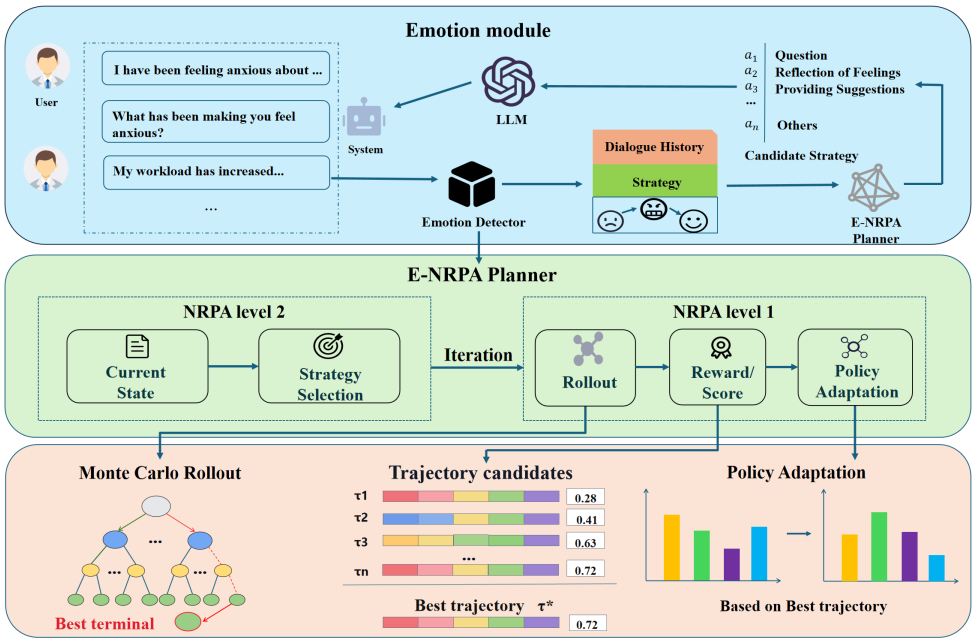


Figure 1: The framework integrates emotion trajectory modeling with a two-level NRPA planner for hierarchical strategy optimization. Level 2 performs high-level strategy selection, while Level 1 conducts rollout evaluation and policy adaptation to guide subsequent LLM response generation.

3.2 Emotion-Aware Policy Planning

To achieve more context-sensitive dialogue planning, E-NRPA incorporates an emotion-aware mechanism into the existing NRPA-GD framework. E-NRPA treats user emotion as part of the planning context. After each user response, a frozen large language model is used to identify the current emotional state, and the detected emotions are maintained as an emotion trajectory. Specifically, the emotion-aware state at turn t is defined as:

$$s_t = (H_t, E_t),$$

where H_t denotes the dialogue history up to turn t, and $E_t = \{e_1, e_2, \ldots, e_t\}$ denotes the historical emotion trajectory, with e_t representing the current user emotion.

3.3 E-NRPA

Figure 1 illustrates the framework of E-NRPA. The core process of E-NRPA centers on nested online simulations and policy adaptation, incorporating the user's emotional trajectory to perform dynamic policy planning during each round of real-world dialogue. At the innermost level (Level 1), the planner performs multiple Monte Carlo simulations based on the current state and strategy to generate complete dialogue trajectories, and calculates a score using a reward function at the end. Once the inner-level simulations are complete, the algorithm selects the simulation trajectory with the highest score and uses NRPA's strategy adaptation mechanism to update the action probability distribution, thereby increasing the weight of strategies that yield high returns under emotion-aware planning contexts. The updated policy is then passed to the outer layer (Level 2) for deeper nested optimization. Through multi-level nested search, E-NRPA can progressively refine the optimal dialogue strategy for the current state, moving from coarse-grained to fine-grained levels.

3.4 Evaluation Tasks

In terms of experiments, this paper examines the performance of E-NRPA on both non-collaborative and collaborative tasks. The non-collaborative tasks include the P4G and CB datasets. The P4G

Method	Backbone	CraigslistBargain			ESConv	
		AT ↓	SR ↑	SL ↑	AT ↓	SR ↑
DialoGPT Zhang et al. (2020)	GPT-2	6.73	0.3245	0.2012	5.31	0.7538
Standard Zhang et al. (2023)	-	6.47	0.3830	0.1588	5.10	0.7692
AnE	-	5.91	0.4521	0.2608	4.76	0.8000
Proactive Deng et al. (2023a)	-	5.80	0.5638	0.2489	5.08	0.7538
+ MI-Prompt Deng et al. (2023b)	-	5.74	0.5691	0.2680	4.78	0.7846
ProCoT Deng et al. (2023a)	-	6.22	0.5319	0.2486	4.75	0.7923
+ MI-Prompt Deng et al. (2023b)	-	6.12	0.5532	0.3059	4.83	0.7769
ICL-AIF Fu et al. (2023)	-	6.53	0.3617	0.1881	4.69	0.8079
PPDPP Deng et al. (2023b)	Vicuna 13B	5.62	0.6117	0.3376	4.56	0.8462
-w/o SFT		5.71	0.6223	0.3354	4.68	0.8384
-w/o RL		5.57	0.6649	0.2280	5.24	0.7308
DPDP (System 1) He et al. (2024)	GPT-3.5-Turbo	5.03	0.7447	0.4108	3.61	0.9000
-System 1 w/o PT		–	–	–	4.22	0.8769
-System 1 w/o SPT		–	–	–	3.97	0.8692
-System 2		2.78	0.9734	0.2728	2.13	0.9923
-System 1 & 2		–	–	–	**2.13**	0.9923
UDP He et al. (2025b)	GPT-4o mini	–	–	–	7.59	0.8320
-w/o PT		–	–	–	7.48	0.7720
-w/o RL		–	–	–	8.64	0.5310
DialogXpert Rakib et al. (2026)	Vicuna 13B	2.93	0.9415	0.3811	2.7	0.9651
-w/o RL		5.13	0.7561	0.3473	4.13	0.8749
DialogXpert	Qwen 1.8B	2.78	0.9274	0.3791	2.49	0.9805
-w/o RL		4.69	0.7754	0.3012	4.04	0.8921
DialogXpert	Qwen2.5 14B	**2.32**	0.9746	0.4389	2.31	0.9876
-w/o RL		3.64	0.8754	0.2952	3.53	0.9401
-w/o LLM-Prior		3.31	0.9165	0.3598	3.89	0.9243
-w/o Emotion		2.75	0.9136	0.3156	3.08	0.9611
NRPA-GD level 1 Wang et al. (2026)	GPT-4o mini	3.89	0.9894	0.6326	5.28	0.9461
NRPA-GD level 2		2.72	**1.0000**	0.6422	4.17	**1.0000**
E-NRPA level 1 (ours)	GPT-4o mini	4.42	0.9716	0.8653	4.26	**1.0000**
E-NRPA level 2 (ours)		4.29	**0.9947**	**0.9345**	4.03	**1.0000**
E-NRPA level 1 (ours)	Qwen2.5 14B	3.13	**1.0000**	0.6842	4.05	**1.0000**
E-NRPA level 2 (ours)		3.34	**1.0000**	**0.7802**	3.64	**1.0000**

Table 1: Comparison of dialogue planning methods on the CraigslistBargain, ESConv benchmarks.

dataset is a persuasive dialogue dataset focused on social welfare scenarios. The CB dataset involves price negotiations and is used to simulate real-world second-hand transaction scenarios He et al. (2018). For collaborative tasks, this study examines the ESConv and ExTES datasets. The ESConv dataset aims to help systems alleviate or resolve users' psychological issues through multi-turn dialogues, while the ExTES dataset serves as an extension to validate E-NRPA's generalization capabilities in more diverse and challenging emotional scenarios Zheng et al. (2024).

3.5 Evaluation Metrics

This paper evaluates the performance of dialogue systems primarily through three core metrics. We first use the average number of dialogue turns (AT) and success rate (SR) He et al. (2024) to measure task completion from the perspectives of efficiency and effectiveness, respectively; SR is defined as

Method	Backbone	P4G		ExTES	
		AT ↓	SR ↑	AT ↓	SR ↑
Standard	-	8.32	0.468	–	–
ProCoT Deng et al. (2023a)	-	7.98	0.543	–	–
ICL-AIF Fu et al. (2023)	-	8.06	0.465	7.65	0.555
GDP-Zero Yu et al. (2023)	-	9.12	0.328	–	–
TRIP Zhang et al. (2024b)	GPT3.5	8.20	0.495	–	–
PPDPP Deng et al. (2023b)	Vicuna 13B	8.185	0.463	8.163	0.558
UDP He et al. (2025b)	GPT-4o mini	7.705	0.598	–	–
- w/o PT		8.017	0.513	–	–
- w/o RL		8.000	0.533	–	–
LDPP He et al. (2025a)	Qwen1-1.8B	5.57	0.795	4.132	0.903
- w/o 2nd Stage		6.14	0.760	4.483	0.865
- w/o 3rd Stage		6.84	0.570	7.038	0.623
DialogXpert Rakib et al. (2026)	Vicuna 13B	5.07	0.8132	2.97	0.9534
DialogXpert	Qwen1-1.8B	3.97	0.8793	2.73	0.9651
DialogXpert	Qwen2.5 14B	3.34	0.9129	**2.57**	0.9782
NRPA-GD level 1 Wang et al. (2026)	GPT-4o mini	3.08	1.0000	4.66	0.9900
NRPA-GD level 2		2.91	1.0000	4.22	1.0000
E-NRPA level 1 (ours)	GPT-4o mini	2.85	**1.0000**	3.94	**1.0000**
E-NRPA level 2 (ours)		2.77	**1.0000**	3.84	**1.0000**
E-NRPA level 1 (ours)	Qwen2.5 14B	1.48	1.0000	4.03	**1.0000**
E-NRPA level 2 (ours)		**1.34**	**1.0000**	3.79	**1.0000**

Table 2: Comparison of dialogue planning methods on the P4G, ExTES benchmarks.

the proportion of cases in which the goal is achieved within a predetermined number of turns. For transaction scenarios, we introduce the Sale-to-List Ratio (SL) to measure the buyer's benefit in price negotiation. Following prior work, SL is computed as a normalized buyer-side benefit score, where a higher value indicates that the final deal is closer to the buyer's target price and therefore more favorable to the buyer. Failed transactions are assigned an SL value of 0. Additionally, to reflect task status in greater detail, this paper introduces a soft success rate (SSR) He et al. (2025a) based on the reward value in the final dialogue turn, serving as a supplement to the traditional SR metric.

3.6 BASELINES

In the experiments of this paper, we select several current mainstream dialogue system methods as baselines to evaluate the effectiveness and superiority of our proposed E-NRPA method. First, we consider methods based on prompt engineering and its enhanced variants: in addition to the standard baseline, we also introduce AnE and ProCoT Deng et al. (2023a), which are representative methods for expert-consultation mechanisms and proactive chain-of-thought planning, respectively; We also include UDP He et al. (2025b) for user-tailored policy planning and ICL-AIF Fu et al. (2023), which improves negotiation through self-play and AI feedback. Second, beyond prompt-based methods, we consider pretrained models and hybrid decision-making frameworks. Specifically, we include DialoGPT Zhang et al. (2020) as a representative pretrained dialogue model. In addition, we compare E-NRPA with PPDPP Deng et al. (2023b), DPDP He et al. (2024), and DialogXpert Rakib et al. (2026), which represent plug-and-play policy planning, dual-process dialogue planning, and online value-based reinforcement learning with LLM priors, respectively. Finally, through a direct comparison with NRPA-GD Wang et al. (2026), we aim to quantify the improvements in online planning performance and interaction quality achieved by the emotional trajectory tracking mechanism when handling complex emotional dynamics.

4 EXPERIMENTS AND ANALYSIS

4.1 MAIN RESULT

We conducted a systematic evaluation of our proposed E-NRPA method on dialogue planning benchmark tasks. The experimental results demonstrate that E-NRPA achieves consistent performance in both cooperative dialogue scenarios (ESConv, ExTES) and non-cooperative dialogue scenarios (CB, P4G). We used GPT-4o-mini and Qwen2.5-14B as backbone models, respectively, to further investigate the method's effectiveness under different model capabilities.

Table 1 presents the results of E-NRPA on the CB and ESConv datasets. On CB, with Qwen2.5-14B as the backbone, E-NRPA Level 2 achieves an AT of 3.34, an SR of 1.0000, and an SL of 0.7802. Although it requires slightly more dialogue turns than DialogXpert, it performs better in both success rate and SL. In particular, its SL improves by 77.76%, showing that E-NRPA can achieve more favorable negotiation outcomes while maintaining reliable task completion. Under the GPT-4o-mini setting, E-NRPA also remains competitive. Compared with NRPA-GD, which does not explicitly model emotion trajectories, E-NRPA shows a slight disadvantage in dialogue turns and success rate but achieves a clear improvement in SL, suggesting that emotion information helps the planner select more advantageous negotiation strategies. On ESConv, E-NRPA consistently reaches an SR of 1.0000, demonstrating stable performance in emotional support tasks. Although it uses slightly more dialogue turns, this may indicate that the model provides more complete support rather than ending the interaction too early.

Table 2 presents the experimental results of E-NRPA on the P4G and ExTES tasks. Under the GPT-4o-mini setting, E-NRPA further reduces the average number of dialogue turns compared with NRPA-GD while maintaining a consistently high success rate. Under the Qwen2.5-14B setting, DialogXpert achieves an AT of 3.34 and an SR of 0.9129 on P4G, whereas E-NRPA Level 2 achieves an AT of 1.34 and an SR of 1.0000, demonstrating greater dialogue efficiency and a higher success rate. On ExTES, DialogXpert achieves an AT of 2.57 and an SR of 0.9782, while E-NRPA Level 2 achieves an AT of 3.79 and an SR of 1.0000. Although E-NRPA requires slightly more dialogue turns on average, it achieves a higher success rate, demonstrating more reliable planning in complex emotional support scenarios. Overall, E-NRPA maintains a consistently high success rate across both tasks, suggesting that incorporating emotional trajectories contributes to more reliable planning and guides dialogues more effectively toward task completion.

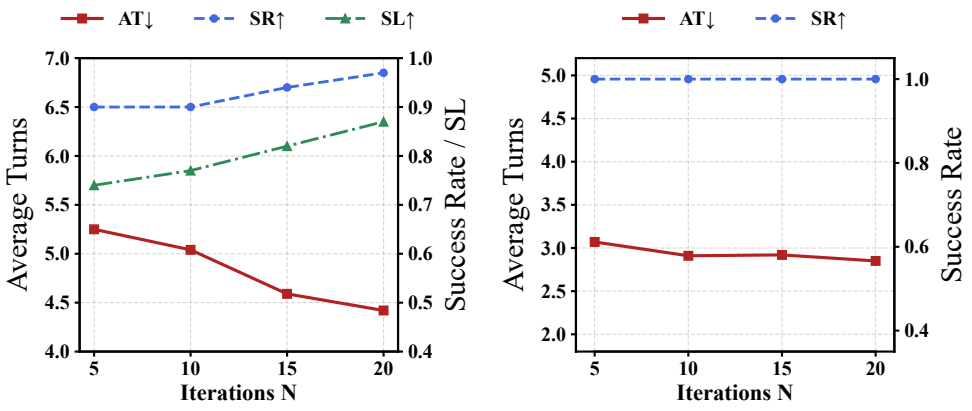
Figure 3 presents the SSR performance of E-NRPA on ESConv and P4G. On ESConv, E-NRPA achieves SSR scores of 0.793 with GPT-4o-mini and 0.824 with Qwen2.5-14B, outperforming all baselines. On P4G, the two versions achieve 0.984 and 0.987, respectively, exceeding NRPA-GD at 0.961 and LDPP at 0.733. These results indicate that emotion perception improves conversational strategy selection and demonstrate the effectiveness of E-NRPA across both emotional support and persuasion-oriented dialogue tasks.

4.2 ABLATION STUDY

Analysis of Iteration N To investigate the effect of NRPA iterations, we evaluated E-NRPA at different values of N, with Figure 2 showing that more iterations generally improve planning quality. On CraigslistBargain, as N increased from 5 to 20, the average dialogue rounds steadily decreased from 5.25 to 4.42, indicating that additional simulations help identify more efficient bargaining strategies. Meanwhile, both SR and SL show consistent upward trends. In particular, SL increases from 0.74 to 0.87, suggesting that additional strategy adaptation improves both task completion and negotiation outcomes.

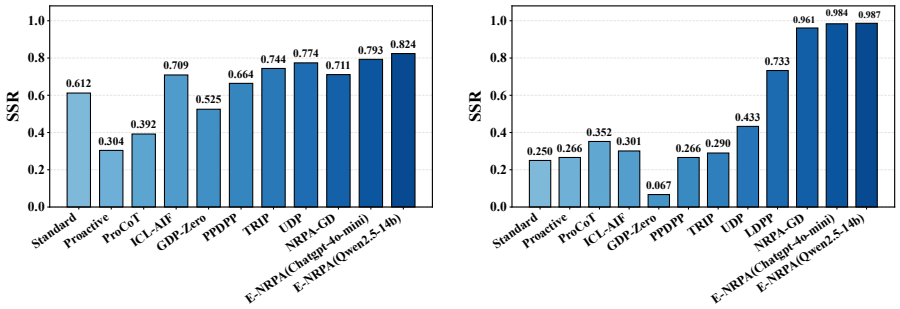
On P4G, the success rate remains 1.00 across all iteration settings, demonstrating reliable task completion even with few simulations. As N increases, the average dialogue turns decrease from 3.07 to 2.85, indicating improved efficiency. Overall, larger iteration budgets help E-NRPA refine action strategies and reduce unnecessary turns while achieving the task goal.

For cooperative tasks, as shown in Table 3, we conducted experiments with different numbers of iterations N on the ESConv and ExTES tasks. Similar to non-cooperative tasks, the number of iterations N has a positive impact on model performance. On the ESConv task, E-NRPA achieves an SR of 1.0000 as early as N=5, indicating that the model can stably complete the task with a



(a) CraigslistBargain (b) P4G

Figure 2: Performance changes under different NRPA iterations.



(a) ESConv **(b) P4G**

Figure 3: Performance comparison of different methods on ESConv and P4G datasets.

relatively small number of simulations. As N increases, the AT gradually decreases from 4.95 to 4.26, suggesting that more iterations help the model find more efficient dialogue strategies. As an extension of the ESConv dataset, ExTES exhibits a similar trend: the success rate (SR) remains at 1.0000 across all settings, while the average number of dialogue turns (AT) decreases from 4.65 to 3.94. Overall, increasing the number of iterations maintains a high success rate while reducing the average number of dialogue turns, indicating that E-NRPA possesses stable planning capabilities in collaborative dialogue tasks and that dialogue efficiency can be further improved through additional simulations.

4.3 Human Evaluation

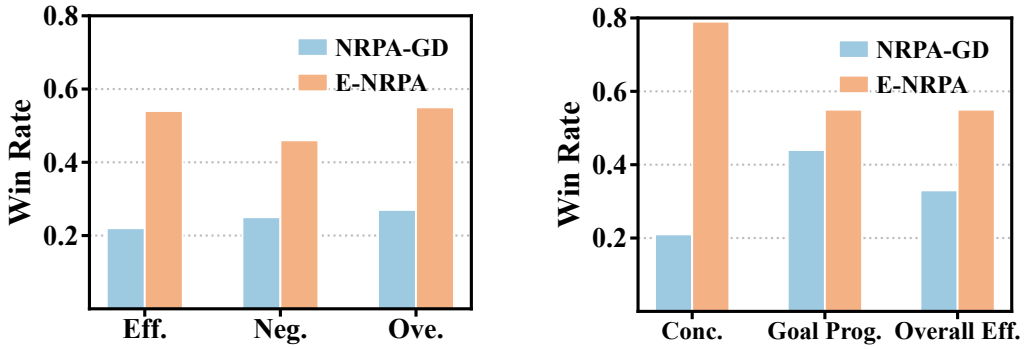
We conducted a human evaluation to compare the dialogues generated by E-NRPA and NRPA-GD. Following prior work Wang et al. (2026), we randomly selected 50 test samples from ESConv and CraigslistBargain (CB). Three annotators independently evaluated anonymized dialogue pairs, yielding 150 judgments for each evaluation dimension.

For ESConv, we evaluated Conciseness (Conc.), Goal Progress (Goal Prog.), and Overall Efficiency (Overall Eff.), which measure redundancy, progress toward resolving user concerns, and the efficiency of emotional support, respectively. For CB, we evaluated Effectiveness (Eff.), Negotiation Ability (Neg.), and Overall Performance (Ove.).

As shown in Figure 4, E-NRPA performs better than NRPA-GD on all three ESConv dimensions. E-NRPA receives preference rates of 79.3%, 54.7%, and 55.3% for Conciseness, Goal Progress, and Overall Efficiency, respectively. The corresponding results for NRPA-GD are 20.7%, 44.0%, and 32.7%. The difference is most pronounced for Conciseness, suggesting that E-NRPA produces more

Method	ESConv		ExTES	
	AT ↓	SR ↑	AT ↓	SR ↑
E-NRPA(N= 5)	4.95	**1.0000**	4.65	**1.0000**
E-NRPA(N=10)	4.68	**1.0000**	4.38	0.9983
E-NRPA(N=15)	4.31	**1.0000**	4.21	**1.0000**
E-NRPA(N=20)	4.26	**1.0000**	**3.94**	**1.0000**

Table 3: Performance of collaborative tasks at different iteration N in the level 1 of E-NRPA.



(a) CraigslistBargain

(b) ESConv

Figure 4: Win rate comparison between NRPA-GD and E-NRPA.

focused responses with less redundant content. Its results on Goal Progress and Overall Efficiency also indicate that these shorter interactions remain effective in advancing the emotional-support process.

E-NRPA also performs well on CB, receiving preference rates of 54.0%, 46.0%, and 55.3% for Effectiveness, Negotiation ability, and Overall performance, respectively. In comparison, NRPA-GD receives 22.0%, 25.3%, and 27.3%. These results suggest that E-NRPA is better able to advance price negotiations and achieve favorable outcomes. This observation is consistent with the automatic evaluation results in Table 1, where E-NRPA achieves a higher SL score.

5 Conclusion

This paper proposes E-NRPA, an emotion-aware dialogue planning framework that integrates user emotion trajectories into nested Monte Carlo simulations and policy adaptation. By evaluating the user's emotional state during the dialogue and incorporating emotion trajectories into nested simulations and policy adaptation, E-NRPA is able to adjust dialogue action selection in real time. Experimental results on cooperative and non-cooperative dialogue tasks demonstrate that E-NRPA achieves strong performance across multiple benchmark datasets, particularly in terms of success rate and dialogue efficiency. Analysis of different numbers of iterations further indicates that additional simulations can improve planning quality while maintaining a stable task success rate. These results suggest that incorporating emotional information into online dialogue planning is an effective way to enhance the adaptability and reliability of dialogue systems. In future work, we plan to further improve the computational efficiency of E-NRPA and explore its application in more complex dialogue scenarios.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and to improve the clarity and readability of the manuscript. All AI-assisted revisions were reviewed and verified by the authors. The authors take full responsibility for the final content of this paper.

REFERENCES

Atheer Algherairy and Moataz Ahmed. A review of dialogue systems: Current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351, 2024.

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

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. Plug-and-play policy planner for large language model powered dialogue agents. *arXiv preprint arXiv:2311.00262*, 2023b.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 807–818, 2024.

Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. Proactive conversational AI: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3):1–45, 2025.

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

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

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

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Yiheng Sun, Zerui Chen, Ming Liu, and Bing Qin. Simulation-free hierarchical latent policy planning for proactive dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24032–24040, 2025a.

Tao He, Lizi Liao, Ming Liu, and Bing Qin. Simulating before planning: Constructing intrinsic user world model for user-tailored dialogue policy planning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 645–655, 2025b.

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

Tazeek Bin Abdur Rakib, Ambuj Mehrish, Lay-Ki Soon, Wern Han Lim, and Soujanya Poria. DialogXpert: Driving intelligent and emotion-aware conversations through online value-based reinforcement learning with LLM priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29967–29975, 2026.

Christopher D. Rosin. Nested rollout policy adaptation for monte carlo tree search. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 649–654, 2011.

Chenwei Wan, Matthieu Labeau, and Chloé Clavel. EmoDynamiX: Emotional support dialogue strategy prediction by modelling mixed emotions and discourse dynamics. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1678–1695, 2025.

Hui Wang, Fafa Zhang, Xiaoyu Zhang, and Chaoxu Mu. A general highly accurate online planning method integrating large language models into nested rollout policy adaptation for dialogue tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33503–33511, 2026.

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

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

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024a.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6665–6694, 2023.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 424–444, 2024b.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

Zhi Zheng, Zhuoliang Xie, Zhenkun Wang, and Bryan Hooi. Monte carlo tree search for comprehensive exploration in LLM-based automatic heuristic design. *arXiv preprint arXiv:2501.08603*, 2025.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11325–11345, 2024.