DialogXpert: Driving Intelligent and Emotion-Aware Conversations Through Online Value-Based Reinforcement Learning with LLM Priors

Tazeek Bin Abdur Rakib[1]*, Ambuj Mehrish[2], Lay-Ki Soon[1,*], Wern Han Lim[1], Soujanya Poria[3,*]

[1]Monash University, Malaysia
[2]Singapore University of Technology and Design, Singapore
[3]Nanyang Technological University, Singapore
{tazeek.binabdurrakib, soon.layki, lim.wern.han}@monash.edu, ambuj_mehrish@sutd.edu.sg, soujanya_poria@ntu.edu.sg

Abstract

Large-language-model (LLM) agents excel at reactive dialogue but struggle with proactive, goal-driven interactions due to myopic decoding and costly planning. We introduce DIALOGXPERT, which leverages a frozen LLM to propose a small, high-quality set of candidate actions per turn and employs a compact Q-network over fixed BERT embeddings trained via temporal-difference learning to select optimal actions within this reduced space. By tracking the user's emotions, DIALOGXPERT tailors each decision to advance the task while nurturing a genuine, empathetic connection. Across negotiation, emotional support, and tutoring benchmarks, DIALOGXPERT drives conversations to under 3 turns with success rates exceeding 94% and, with a larger LLM prior, pushes success above 97% while markedly improving negotiation outcomes. The proposed framework delivers real-time, strategic, and emotionally intelligent dialogue planning at scale.

1 Introduction

Before the emergence of LLMs, dialogue policy planning largely relied on supervised learning approaches, where classifiers were trained on annotated corpora of dialogue acts and high-level strategies to predict system responses (Zhou et al. 2020; Joshi et al. 2021; Cheng et al. 2022). These models tend to be domain-specific, static, and costly to extend. Moreover, most deployed agents remain reactive: they respond to what the user just said, but rarely steer the conversation toward longer-term objectives. However, scenarios such as negotiation, tutoring, and emotional support demand initiative, sustained strategizing, and empathy (Deng et al. 2023; Kang et al. 2024; Song et al. 2024). While current LLMs have enabled new forms of open-domain interaction, they often fall short in proactive roles, as single-step decoding optimizes for the next token rather than long-term conversational outcomes (Levin, Pieraccini, and Eckert 1997; Cheng et al. 2022).

Recent advances in LLMs, such as ChatGPT (OpenAI 2022), Vicuna (Zheng et al. 2023a), and LLaMA2-Chat (Ouyang et al. 2022; Touvron et al. 2023), have significantly improved open-domain dialogue systems by enabling fluent, context-aware, and intent-aligned responses (Hu et al. 2023). Rather than relying on labeled supervision at every turn, modern planners now query LLMs to simulate plausible conversational outcomes. State-of-the-art (SOTA) approaches include Plug-and-Play Dialogue Policy Planning (PPDPP) (Deng et al. 2024) and search-based methods like Dual-Process Dialogue Planner (DPDP) (He et al. 2024), which incorporate planning modules to reason over multiple turns and optimize decision-making.

PPDPP fine-tunes a RoBERTa (Liu et al. 2019) policy network on self-play dialogues generated by frozen LLM-based user and reward simulators, then selects the single highest-scoring action at each turn. This one-pass architecture is computationally efficient but myopic: it lacks foresight and struggles with states outside its training distribution. DPDP introduces Kahneman's "System-2" thinking: when its fast RoBERTa policy ("System-1") is uncertain, it triggers Monte Carlo Tree Search (MCTS) rollouts guided by LLM-based reward estimates (Silver et al. 2016; Zhao et al. 2024). While deeper lookahead improves task success, repeated simulations drastically increase latency and token cost. Moreover, the DPDP heuristic gating may misfire that results in invoking expensive search unnecessarily or skipping it when crucial. These approaches highlight recurring bottlenecks in current SOTA dialogue planning: MCTS latency impedes real-time deployment, emotion blindness risks tone-deaf replies (Chen et al. 2023; Asghar et al. 2020), and sample-hungry fine-tuning ties performance to costly data collection.

In this paper, we present DIALOGXPERT, addressing key limitations through the LLM-Prior planning paradigm. Instead of expanding an entire search tree, a frozen LLM such as Qwen2.5-14B (Bai et al. 2023) proposes a top-k set of semantically coherent actions for the current state, forming a concise prior over the action space (Bengio 2017; Korbak, Perez, and Buckley 2022). A lightweight Q-network, operating on fixed BERT embeddings (Devlin 2018) and trained via off-policy temporal-difference learning (Mnih et al. 2013; Watkins and Dayan 1992; Tesauro et al. 1995), then estimates the long-term value of each candidate. Since evaluation is restricted to a small candidate set, DIALOGXPERT retains LLM flexibility without the cost of full tree expansion. In addition, DIALOGXPERT is the first proac-

*Corresponding author

tive dialogue planner to explicitly incorporate emotion. After each turn, an emotion tracker infers the user's affective state from context. By guiding LLM priors with emotional cues, DIALOGXPERT avoids abrupt or tone-deaf responses (Zhao et al. 2023). Our main contributions are as follows.

- **Emotion-aware planning**: Introduce a novel integration of emotion trajectories in goal-driven conversations. DIALOGXPERT seamlessly fuses affect tracking with goal-directed LLM planning, enabling emotion-aware selection in goal-driven conversations
- **Action space reduction**: Leverage LLM priors to reduce action space. By leveraging LLM priors to reduce planning complexity via selecting the top-k actions, we are able to both reduce the number of required sample episodes and replace the need for fine-tuning plug-play models that may lead to possible bias.
- **Efficiency against MCTS**. Improve efficiency by combining LLM prior action-space with Deep Q-learning and frozen BERT model. This improves efficiency drastically and achieves comparable performance against MCTS.
- **Extensive validation**: Conduct experiments on five diverse datasets spanning collaborative and non-collaborative dialogue settings. Results demonstrates strong generalization and empirical gains across success rate, average turns, and on human evaluation settings.

2 Related Works

LLM-powered Dialogue Agent. LLM-driven decision-making has progressed from fine-tuned chatbots to sophisticated planners. Early prompt-based systems like DialoGPT (Zhang et al. 2019, 2020), ProAgent (Zhang et al. 2023), and Voyager (Wang et al. 2023) adapted pretrained transformers or retrieval-augmented controllers for multi-step tasks. Subsequent techniques like prompt-chaining (Proactive, ProCoT (Deng et al. 2023)) and modular prompting (Ask-an-Expert (Zhang, Naradowsky, and Miyao 2023), ICL-AIF (Fu et al. 2023)) enabled iterative reasoning and task decomposition. Planner-based search methods, including Tree-of-Thoughts (Yao et al. 2023) and MCTS-based rollouts (Hao et al. 2023), and reinforcement learning approaches like PPDPP (Deng et al. 2024) and DPDP (He et al. 2024) further improved exploration efficiency. Recently, latent-policy techniques such as LDPP (He et al. 2025a) and UDP (He et al. 2025b) learn continuous action representations via VAE and diffusion-based user models. In contrast, DIALOGXPERT treats the LLM as a frozen action proposer selecting top-k actions aligned with both the conversational context and user emotional trajectory. These actions are then ranked with via Q-learning enabling fast, strategic, and emotionally grounded decisions—without requiring full-tree search at inference time.

Emotions in conversations. While emotion recognition in conversations remains an active area of research (Kang and Cho 2025; Li et al. 2025), its integration into dialogue systems has largely focused on different contexts, specifically response generation (Asghar et al. 2020) and dialogue state tracking (DST) (Balaraman, Sheikhalishahi, and Magnini 2021). In response generation, emotions are used to improve empathy and alignment at the utterance level, through methods ranging from word-level emotion control to affective therapy-inspired prompting (Rashkin et al. 2018). However, these systems are often limited to single-turn settings (Chen et al. 2023) and do not leverage emotion for multi-turn, goal-oriented strategy planning. In DST, the focus is on intent tracking to support reactive responses, typically using task-oriented datasets like MultiWOZ(Budzianowski et al. 2018), which lack emotional signals. Consequently, DST models prioritize information retrieval over user engagement or conversational flow. In contrast, our work centers on proactive, goal-driven dialogue planning by incorporating emotion trajectories directly into the state space. These trajectories are tracked over the course of the conversation and used to guide top-k action selection via LLM priors (Yan et al. 2024), enabling more emotionally aligned and strategic planning. This use of emotion trajectories represents a key departure from prior work focused solely on empathetic response generation.

Reducing action space. There are several methods for reducing the action space in reinforcement learning. Two common approaches are action masking (Stolz et al. 2024) and imitation learning (Zare et al. 2024). Action masking eliminates invalid or infeasible actions during training or inference (Liang et al. 2023). It is typically used in structured environments (Varricchione et al. 2024), such as games, where valid actions are clearly defined by rules (Hou et al. 2023). Imitation learning, on the other hand, relies on expert-labeled demonstrations to "warm start" a model through supervised pretraining. For instance, models like PPDPP and DPDP fine-tune RoBERTa-based planners using imitation learning. However, both action masking and imitation learning face limitations in open-ended, goal-driven conversations. First, defining reliable heuristics for masking actions is infeasible due to the context-sensitive and diverse nature of conversations. Second, imitation learning is sensitive to data imbalance and annotation bias, which may limit generalization. To address these challenges, our method avoids both heuristics and fine-tuning by leveraging LLM priors. Specifically, we use a frozen LLM to generate a rational top-k subset of actions, which serves as the candidate pool for DQN-based value estimation.

3 Methodology

3.1 Preliminaries

Problem statement. Existing works (Wang et al. 2020; He et al. 2024, 2025a) formulate the dialogue planning process as a Markov Decision Process (MDP), represented formally as a tuple $(\mathcal{S}, \mathcal{A}, r, \mathcal{T})$, where $\mathcal{S}$ denotes the dialogue state space, $\mathcal{A}$ represents the dialogue action space, r denotes the reward function, and $\mathcal{T}$ defines the transition function. At each turn t, the dialogue state $s_t \in \mathcal{S}$ includes the complete conversational context and encompassing historical utterances. The agent selects an action $a_t \in \mathcal{A}$, which leads to a state transition $s_{t+1} = \mathcal{T}(s_t, a_t)$ and a reward r_t. The goal of the dialogue agent is to learn an optimal policy π^* maximizing cumulative future rewards:

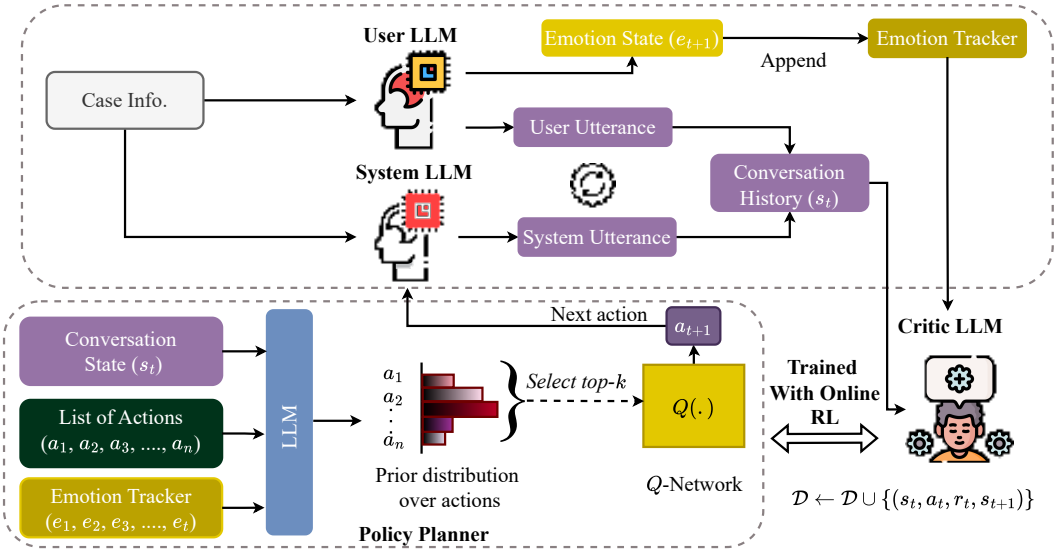

Figure 1: DIALOGXPERT pipeline: case information and dialogue history drive user/system LLMs and an emotion tracker; a frozen LLM generates a prior over candidate actions, the top-k are evaluated by a Q-network and executed by the system LLM; a critic LLM provides reward signals to train the Q-network.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T} \gamma^t r_t \right] \tag{1}$$

where $\gamma \in [0, 1]$ is the discount factor and T is the maximum dialogue length.

LLM-powered self-play. Following (He et al. 2024, 2025a), we leverage LLMs to simulate both user and system roles in order to generate realistic self-play dialogues. LLMs are utilized for their strong token priors, consistent role-conditioned generation, and calibrated scoring—capabilities essential for stable self-play and effective reward shaping. Two distinct LLM agents are used: one represents the user and the other the dialogue system, as illustrated in Figure 1. Given predefined case information (Case Info.), each agent generates utterances conditioned on its role and the prior conversation history (Luo et al. 2022). Additionally, an independent LLM-based critic evaluates each turn, assigning scalar rewards that reflect both task success and emotional alignment, thereby enabling reinforcement learning. Further implementation details are provided in Appendix C.

Following PPDPP and DPDP, we use the same frozen LLM across all three roles: User, System, and Critic. PPDPP (Deng et al. 2024) showed that a mid-sized Vicuna-13B achieves over 90% F1-score when serving as both the reward model and user simulator on CB and ESConv (Liu et al. 2021), demonstrating that such models can function as reliable evaluators and interlocutors without task-specific fine-tuning. Sharing a single backbone across roles offers several advantages: (i) it eliminates confounds, allowing performance gains to be attributed solely to the planner; (ii) it keeps the full pipeline within memory constraints; and (iii) it ensures consistent natural-language feedback, which is converted into scalar rewards for Q-network training (see Appendix E).

3.2 LLM Action Prior Framework

The LLM Action Prior Framework leverages the semantic knowledge of pretrained LLMs to narrow the dialogue action space. By conditioning on the current dialogue state s_t including conversational history and emotional context—the LLM generates a prior distribution over candidate actions, significantly reducing computational overhead and guiding effective action selection. Formally, this prior is defined as $p_{\text{LLM}}(\cdot \mid s_t)$.

Following (Yan et al. 2024), we adopt a two-step "free-form + projection" approach that combines the generative flexibility of LLMs with a constrained action space $\mathcal{A} = \{a_1, \ldots, a_n\}$. At each dialogue turn t, the model input is: $\mathcal{I} = (c_t, s_t, E_t)$,, where c_t is the case information, s_t includes the conversation history, and E_t represents the accumulated emotion. The input $\mathcal{I}$ and action set $\mathcal{A}$ are serialized into a prompt (see Appendix A). The LLM produces an open-text proposal:

$$o \sim p_{\text{LLM}}(o \mid s_t, \mathcal{A}) \tag{2}$$

which is projected via a deterministic mapping $\mathcal{P}$ to a valid action: $a_{t+1} = \mathcal{P}(o) \in \mathcal{A}$. Comprehensive technical details are provided in (Ye et al. 2024), and our specific adaptation is described in Appendix A.

Although we do not enumerate the full action space internally, including $\mathcal{A}$ in the prompt implicitly defines a normalized prior over actions, denoted $p_{\text{proj}}(a \mid s_t)$. From this distribution, we extract the top-k most probable actions:

$$A_t^{\text{top-}k} = \text{Top-}k(p_{\text{proj}}(a \mid s_t)) \tag{3}$$

This approach reduces the dimensionality and complexity of decision-making by focusing computation on a compact set of semantically coherent, contextually appropriate candidate actions.

Q-Network: In our implementation (illustrated in Figure 1), the action-value function $Q(s, a)$ uses a pretrained BERT encoder[1] (kept fixed) followed by a lightweight adaptor network (3 layer MLP). Specifically, given the current state s_t and each proposed action a_i (sampled via the free-form + projection prior), we construct the input sequence:

`[CLS] State:` s_t `[SEP] Action:` a_i `[SEP]`

tokenize it, and feed it into BERT. We take the final hidden vector $\mathbf{h}_i \in \mathbb{R}^d$ at the `[CLS]` position and pass it through a three-layer MLP adaptor with ReLU activations to produce a scalar score: $\tilde{Q}_i = \mathrm{BERT}_{\mathrm{Adaptor}}(\mathbf{h}_i) \in \mathbb{R}$.

We then normalize these scores across all K candidates using a softmax,

$$p_Q(a_i \mid s_t) = \frac{\exp(\tilde{Q}_i)}{\sum_{j=1}^{K} \exp(\tilde{Q}_j)} \tag{4}$$

and select the highest-probability action $a^* = \arg\max_i p_Q(a_i \mid s_t)$. The chosen a^* is executed to produce the next state. Rather than a purely greedy policy, we adopt an ϵ-greedy strategy with ϵ chosen empirically.

3.3 Emotion-Aware Policy Planning

Integrating emotional context into dialogue policy planning is critical for building proactive, user-aligned systems (Zhao et al. 2023). Unlike earlier planners that rely solely on semantic or task-specific signals (Wang et al. 2020), our approach introduces an *Emotion Tracker* that queries a frozen LLM at every user turn t to produce a lightweight, auxiliary affect label e_t, without requiring additional embeddings or fine-tuning.

$$e_t = \text{LLM-EmoPred}(u_t^{\text{usr}}) \tag{5}$$

The accumulated emotional trajectory E_t is defined as the sequence of predicted emotional states up to turn t, where $E_t = [e_1, e_2, ..., e_t]$. This trajectory is fused with the dialogue context to form an emotion-aware state, as illustrated in Figure 1, where s_t denotes the conversation state. Both the LLM prior, which proposes the top-k candidate actions, and the BERT-based Q-network, which ranks them, operate on this enriched sub-actions state, ensuring that generation and selection are explicitly emotion-conditioned. Additionally, E_t is forwarded to the Critic LLM so that reward signals jointly reflect task progress and affective alignment, guiding policy learning toward strategies that are both effective and empathetic. Further information is given in Appendix C.

3.4 Online RL with LLM Priors

At each dialogue turn t, we first query the free-form + projection LLM prior to obtain a distribution $p_{\mathrm{proj}}(a \mid s_t)$ over the finite action set $\mathcal{A}$. Rather than sampling directly from this prior, we evaluate each candidate action $a \in \mathcal{A}$ with Q-network and select the action with the highest value:

$$a_t = \arg\max_{a \in \mathcal{A}} Q^\theta(s_t, a) \tag{6}$$

We then execute a_t in the environment, observe the next state s_{t+1}, and solicit a scalar reward r_t from the Critic LLM, which assesses the transition (s_t, a_t, s_{t+1}) in terms of task effectiveness and emotional alignment. The tuple (s_t, a_t, r_t, s_{t+1}) is appended to the replay buffer $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r_t, s_{t+1})\}$.

Periodically, we sample minibatches from $\mathcal{D}$ and use TD-learning to update Q-network. The loss function is given as:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \Big[Q^\theta(s, a) - y \Big]^2. \tag{7}$$

where $y = r_t + \gamma \max_{a' \in \mathcal{A}} Q^\theta(s_{t+1}, a')$. Compared to traditional DQN, the DQN-Prior performs exploration and applies the Bellman optimal operator within the LLM prior action space $\mathcal{A}$. Throughout training, all exploratory actions and Bellman backups draw from the LLM-induced prior, while the Critic LLM's rewards guide the Q-network toward semantically coherent and emotionally aware dialogue policies.

4 Experimental Setup

Datasets and Reward Values We evaluate our method on five proactive dialogue datasets spanning diverse domains. ESConv (Liu et al. 2021) focuses on emotional support, CIMA (Stasaski, Kao, and Hearst 2020) features English-to-Italian tutoring dialogues, CraigslistBargain (CB) (He et al. 2018) involves buyer-seller negotiations, P4G (Wang et al. 2019) covers persuasive donation dialogues, and ExTES (Zheng et al. 2023b) extends ESConv to more varied emotional contexts. Datasets are grouped into **collaborative** (ESConv, CIMA, ExTES) and **non-collaborative** (CB, P4G) settings based on whether participants share a common goal.

Predefined action prompts (Appendix G.5) and case backgrounds are used to initialize dialogue states. An LLM-based critic provides scalar rewards tailored to each task domain, with full reward structure and scoring heuristics detailed in Appendix E. Dataset splits and further background information are provided in Appendix D, and full implementation details are in Appendix C.

Baselines In addition to DialoGPT (Zhang et al. 2019), we compare DIALOGXPERT against two groups of models: prompt-based and planner-based dialogue models. Prompt-based models include Standard Prompting, Proactive and ProCoT (Deng et al. 2023), Ask-an-Expert (Zhang, Naradowsky, and Miyao 2023), and ICL-AIF (Fu et al. 2023). Planner-based approaches represent the current SOTA approaches including PPDPP (Deng et al. 2024), DPDP (He et al. 2024), LDPP (He et al. 2025a), and UDP (He et al. 2025b). All of the planner-based pproaches are centered on RoBERTa as the core planner. The results of prompt-based models are reported from (Deng et al. 2024) and the results

[1]https://huggingface.co/google-bert/bert-base-uncased

Method	Backbone	CraigslistBargain			ESConv		CIMA	
		AT ↓	SR ↑	SL ↑	AT ↓	SR ↑	AT ↓	SR ↑
DialoGPT (Zhang et al. 2019)	GPT-2	6.73	0.3245	0.2012	5.31	0.7538	5.43	0.4956
Standard	-	6.47	0.3830	0.1588	5.10	0.7692	3.89	0.6903
AnE (Zhang, Naradowsky, and Miyao 2023)	-	5.91	0.4521	0.2608	4.76	0.8000	3.86	0.6549
Proactive (Deng et al. 2023)	-	5.80	0.5638	0.2489	5.08	0.7538	4.84	0.5310
+ MI-Prompt (Deng et al. 2024)	-	5.74	0.5691	0.2680	4.78	0.7846	4.70	0.5664
ProCoT (Deng et al. 2023)	-	6.22	0.5319	0.2486	4.75	0.7923	4.58	0.5487
+ MI-Prompt (Deng et al. 2024)	-	6.12	0.5532	0.3059	4.83	0.7769	4.72	0.5221
ICL-AIF (Fu et al. 2023)	-	6.53	0.3617	0.1881	4.69	0.8079	4.19	0.6106
PPDPP (Deng et al. 2024)	Vicuna 13B	5.62	0.6117	0.3376	4.56	0.8462	3.03	0.8407
-w/o SFT		5.71	0.6223	0.3354	4.68	0.8384	3.18	0.8230
-w/o RL		5.57	0.6649	0.2280	5.24	0.7308	3.41	0.7965
DPDP (System 1) (He et al. 2024)	GPT-3.5-Turbo	5.03	0.7447	0.4108	3.61	0.9000	2.24	0.9469
-System 1 w/o PT		–	–	–	4.22	0.8769	2.36	0.9292
-System 1 w/o SPT		–	–	–	3.97	0.8692	2.51	0.8938
-System 2		2.78	0.9734	0.2728	2.13	0.9923	2.49	0.9735
-System 1 & 2		–	–	–	**2.13**	**0.9923**	2.28	0.9823
UDP (He et al. 2025a)	GPT-4o mini	–	–	–	7.59	0.8320	–	–
-w/o PT		–	–	–	7.48	0.7720	–	–
-w/o RL		–	–	–	8.64	0.5310	–	–
DialogXpert	Vicuna 13B	2.93	0.9415	0.3811	2.70	0.9651	2.24	0.9883
-w/o RL		5.13	0.7561	0.3473	4.13	0.8749	3.05	0.8829
DialogXpert	Qwen 1.8B	2.78	0.9274	0.3791	2.49	0.9805	2.16	0.9902
-w/o RL		4.69	0.7754	0.3012	4.04	0.8921	2.96	0.9042
DialogXpert	Qwen2.5 14B	**2.32**	**0.9746**	**0.4389**	2.31	0.9876	**2.03**	**0.9951**
-w/o RL		3.64	0.8754	0.2952	3.53	0.9401	2.62	0.9317
-w/o LLM-Prior		3.31	0.9165	0.3598	3.89	0.9243	2.71	0.9395
-w/o Emotion		2.75	0.9136	0.3156	3.08	0.9611	2.34	0.9425

Table 1: Comparison of dialogue planning methods on the CraigslistBargain, ESConv and CIMA benchmarks. The colors indicate same LLM backbone. Best results are shown in **bold** and the second best are underlined.

of planner-based dialogue models are taken from their respective papers. Full model descriptions are shown in Appendix C.

Metrics: Following PPDPP (Deng et al. 2024) and DPDP (He et al. 2024), we evaluate dialogue quality using two primary metrics: **Average Turn (AT)**, which measures conversational efficiency by computing the mean number of turns needed to reach the goal (Kwan et al. 2023); and **Success Rate (SR)**, which reflects the proportion of successful outcomes within a fixed turn limit (Gao et al. 2021). For the CB dataset, we additionally report the **Sale-to-List Ratio (SL)** (Zhou et al. 2019), which captures negotiation quality from the buyer's perspective.

LLM Variations: Several baseline planners incorporate different LLM backbones and planning strategies. For example, DialoGPT uses GPT-2 for greedy turn-by-turn responses while UDP/LDPP exploits GPT-4o-mini or Qwen 1.8B for latent policy mining. In our experiments, DIALOGXPERT treats the LLM as a frozen action proposer, generating a set of top-k action candidates per turn. We use the following LLMs: Vicuna-13B, Qwen1-1.8B, and Qwen2.5 14B to analyze how different LLM backbones affect planning behavior, action diversity, and emotional alignment.

5 Results and Analysis

5.1 Main Results

Table 1 summarizes the performance of diverse baselines, MCTS-style planners, recent policy-LM methods, and our two DIALOGXPERT variants (Vicuna 13B and Qwen 2.5 14B). Furthermore, preliminary experiments identified $\epsilon = 0.5$ and top-$k = 4$ as optimal, and these values are fixed in all subsequent evaluations. By integrating an LLM-prior policy with lightweight value learning and emotion tracking, DIALOGXPERT achieves sub-3-turn dialogues and success rates above 94% across all three benchmarks (CB, ESConv, and CIMA) with the Vicuna backbone, and further improves to $SR > 97\%$ and $SL = 0.4389$ with Qwen 2.5 14B for CB, while maintaining average turns around 2.32. As shown in Tables 1 and 3, DIALOGXPERT not only surpasses DPDP and PPDPP in both efficiency and effectiveness, but also generalizes strongly across diverse settings including P4G and ExTES, where it delivers the highest success rates (97.2% on ExTES) and competitive turn efficiency. These results confirm that DIALOGXPERT offers a practical alternative to computationally intensive planning approaches, without sacrificing quality. **We performed paired t-tests against the PPDPP baselines, and all p-values were below**

MCTS Ratio	CraigslistBargain			ESConv		CIMA	
	AT ↓	SR ↑	SL ↑	AT ↓	SR ↑	AT ↓	SR ↑
22.3 % MCTS	3.69	0.8298	0.3102	–	–	–	–
51.4 % MCTS	2.77	0.9468	0.3118	–	–	–	–
60.3 % MCTS	2.49	0.9681	0.2856	–	–	–	–
0.0 % MCTS	–	–	–	3.61	0.9000	–	–
21.9 % MCTS	–	–	–	3.42	0.9154	–	–
46.5 % MCTS	–	–	–	2.95	0.9692	–	–
68.3 % MCTS	–	–	–	2.72	0.9769	–	–
100 % MCTS	–	–	–	**2.13**	**0.9923**	–	–
0.0 % MCTS	–	–	–	–	–	2.24	0.9469
28.6 % MCTS	–	–	–	–	–	2.39	0.9646
50.0 % MCTS	–	–	–	–	–	2.28	0.9823
81.1 % MCTS	–	–	–	–	–	2.58	0.9735
100 % MCTS	–	–	–	–	–	2.49	0.9735
DialogXpert (Vicuna 13B)	2.93	0.9415	0.3811	2.70	0.9651	2.24	0.9883
DialogXpert (Qwen 2.5 14B)	**2.32**	**0.9746**	**0.4389**	2.31	0.9876	**2.03**	**0.9951**

Table 2: Comparisons of MCTS variants (using GPT3.5-Turbo) of DPDP against DIALOGXPERT. Scores are taken from (He et al. 2024) and the numerical values indicate the MCTS ratio. Best results are shown in **bold** and the second best are underlined.

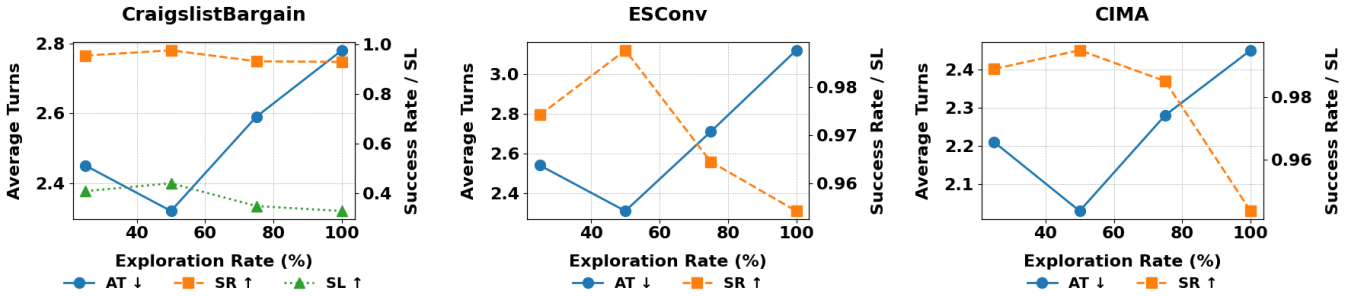

Figure 2: Comparisons of different ϵ values for Exploration vs. Exploitation. We use the Qwen 2.5 14B prior and sweep the ϵ-greedy parameter (ϵ) to measure how different exploration rates impact performance.

the standard 0.05 threshold, indicating statistical significance.

Comparison with MCTS Variants: We compare DPDP's MCTS-based planner with our DIALOGXPERT variants in Table 2. While MCTS rollouts improve DPDP's performance (e.g., reducing AT from 3.69 to 2.49 and lifting SR from roughly 83% to 97% on CB) as the rollout ratio increases, they incur high simulation and latency costs. In contrast, DIALOGXPERT (Vicuna-13B and Qwen2.5-14B) achieves comparable, or better performance, without tree search. On ESConv, DIALOGXPERT achieves an AT of 2.31 turns with 98.76% SR, almost matching DPDP's 100% MCTS rollout performance. More importantly, DIALOGXPERT surpasses 100% MCTS on CIMA and CB for AT, SR, and SL as shown in Table 2.

Cost and Efficiency Analysis: SOTA baselines (PPDPP, DPDP, and LDPP) rely on RoBERTa models with task-specific fine-tuning and offline reinforcement learning. This introduces additional training overhead and potential bias from the fine-tuning dataset. DIALOGXPERT improves on these in three ways. First, it removes the need for pretraining by using a frozen LLM to select the candidate actions, reducing annotation and retraining costs. Second, unlike DPDP's MCTS based approach that requires about 30 LLM calls per turn, DIALOGXPERT reduces this to just 4 by leveraging top-k sampling from the LLM prior to narrow the action space. Finally, this focused decoding is paired with a lightweight DQN for value estimation that enables efficient decision-making without exhaustive simulation. The LLMs and the BERT encoder remain frozen during training, with only the Q-network being updated. This promotes stable and efficient learning via continual policy refinement using diverse state-action pairs from the replay buffer, enabling strong adaptation with minimal training cost.

Human Evaluation We also conducted human evaluations (Joshi et al. 2021; Liu et al. 2021) on the ESConv dataset. Four annotators assessed system responses across four criteria: Suggestion, Identification, Comforting, and Overall Quality. Annotators compared system outputs and labeled each metric as a win, loss, or tie, with final scores averaged across all judgments. To ensure a fair compari-

Method	Backbone	P4G		ExTES	
		AT ↓	SR ↑	AT ↓	SR ↑
Standard	-	8.32	0.468	–	–
ProCoT (Deng et al. 2023)	-	7.975	0.543	–	–
ICL-AIF (Fu et al. 2023)	-	8.085	0.465	7.65	0.555
GDP-Zero (Yu, Chen, and Yu 2023)	-	9.119	0.328	–	–
TRIP (Zhang et al. 2024)	GPT3.5	8.20	0.495	–	–
PPDPP (Deng et al. 2024)	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	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**

Table 3: Evaluation of dialogue planners on P4G and ExTES. Colors indicate same LLM backbone. Best results are shown in **bold**.

son, both DIALOGXPERT and PPDPP were run with the same Vicuna-13B backbone on 20 randomly selected ESConv emotional-support dialogues. As illustrated in Figure 3:,DIALOGXPERT outperforms PPDPP on Identification (60% vs. 35%), Comforting (52% vs. 45%), and Overall Quality (51% vs. 41%), with modest tie rates and lower loss rates. These results highlight that DIALOGXPERT offers a better balance between empathy and actionable support, underscoring the value of incorporating emotional trajectories and selecting coherent actions using LLM priors. Annotator instructions and additional human evaluations on the CIMA dataset are provided in Appendix B.

Generalization Test: Following (He et al. 2025b), we assess generalization from ExTES to ESConv, given their similar environments and action labels (differing only in reward computation). We train the Q-network on ExTES and directly evaluate it on ESConv without further fine-tuning. Our approach achieves AT of 2.28 (vs 5.39) and SR of 99.43% (vs. 78.10%), significantly outperforming LDPP. This strong transfer performance stems from the larger training set in ExTES, enabling better generalization. In contrast, LDPP relies heavily on RoBERTa-based encoders/decoders, making it more sensitive to domain shifts.

5.2 Ablation Studies

Impact of Emotions: We examine the impact of incorporating emotional trajectories. As shown in Table 1, removing the emotion trajectory (denoted as *w/o Emotion* in Table 1) leads to consistent performance drops across all. For example, the AT on CB increases from 2.32 to 2.75 and SR decreases from 97.46% to 91.36%. Moreover, there is a drastic drop of SL from 0.4389 to 0.3156. These differences highlight that incorporating an emotion tracker improves both efficiency and outcome quality. This is because the candidate actions, generated via LLM priors, are more contextually grounded in the user's emotional state. As a result, the Q-network operates over a more focused and emotionally aligned sub-action space. While the Q-network does not directly observe the emotion trajectory, it indirectly benefits by evaluating actions filtered through emotion-informed context. This demonstrates that emotion trajectories are not merely auxiliary, but serve as a critical control signal. Additional ablations regarding emotion trajectories are provided in Appendix F.

Impact of LLM Prior LLM prior narrows the action space to relevant candidates, reducing computation and boosting decision quality. Disabling it causes drop in performance. We can observe in Table 1 that on ESConv, success falls from 98.76% to 94.01% and average turns rise from 2.31 to 3.53; on CIMA, success drops from 99.51% to 93.17%. Without the prior, the agent repeats trivial patterns and struggles to choose optimal actions. By providing diverse, high-quality options, the prior lets the Q-network focus on value learning its removal degrades efficiency, planning, and generalization.

Exploitation vs Exploration As illustrated in Figure 2, our ϵ-greedy strategy controls the trade-off between exploration and exploitation (Tokic 2010). At ϵ=25%, we surpass pure LLM inference ($\tilde{9}5\%$ success, SL = 0.407) but may overlook best actions; at $\epsilon \geq 75\%$, performance dips (turns > 2.5, success < 97%, SL < 0.35); and at $\epsilon = 100\%$, all learned value is ignored. The sweet spot is $\epsilon = 50\%$, yielding the fewest turns (2.32 negotiation, 2.31 support, 2.03 tutoring) with peak success (97.5–99.5%) and SL = 0.439, confirming that moderate exploration maximizes planning efficiency.

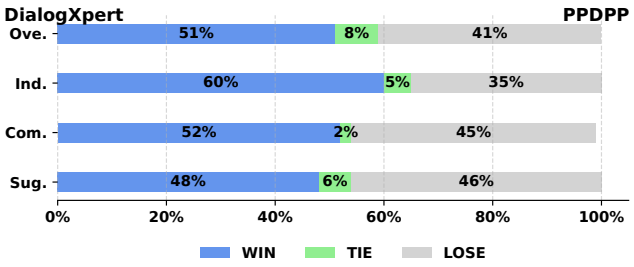

Figure 3: Win/tie/loss percentages for DIALOGXPERT vs. PPDPP on ESConv across Identification, Comforting, Suggestion and Overall metrics.

6 Conclusion and Future Work

We introduced DIALOGXPERT, a novel framework that combines frozen LLM priors, lightweight value-based RL, and emotion tracking to enable proactive and emotionally intelligent dialogue planning. Across negotiation, emotional support, and tutoring tasks, DIALOGXPERT delivers shorter, more effective conversations and higher success rates than both fine-tuned policy LMs and MCTS-based planners. By narrowing the action space through LLM priors and incorporating emotion signals, our model generalizes well across tasks while producing more empathetic, user-aligned dialogues. Looking ahead, dynamic adjustment of the LLM prior could improve adaptability to user feedback. Multimodal integration (e.g., visual or auditory inputs) may further enrich context and interactivity.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-GV-2023-010). This work is also supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-005), and the NTU SUG project #025628-00001:Post-training to Improve Embodied AI Agents.

References

Asghar, N.; Kobyzev, I.; Hoey, J.; Poupart, P.; et al. 2020. Generating emotionally aligned responses in dialogues using affect control theory. *arXiv preprint arXiv:2003.03645*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; et al. 2023. Qwen technical report. *arXiv*.

Balaraman, V.; Sheikhalishahi, S.; and Magnini, B. 2021. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the special interest group on discourse and dialogue*.

Bengio, Y. 2017. The Consciousness Prior. *ArXiv*.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; et al. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv*.

Chen, J.; Yang, S.; Xiong, J.; and Xiong, Y. 2023. An effective emotion tendency perception model in empathic dialogue. *Plos one*.

Cheng, Y.; Liu, W.; Li, W.; Wang, J.; et al. 2022. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. *CoRR*.

Deng, Y.; Lei, W.; Liao, L.; and Chua, T.-S. 2023. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. *arXiv*.

Deng, Y.; Zhang, W.; Lam, W.; Ng, S.-K.; et al. 2024. Plug-and-Play Policy Planner for Large Language Model Powered Dialogue Agents. In *ICLR*.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

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*.

Gao, C.; Lei, W.; He, X.; De Rijke, M.; et al. 2021. Advances and challenges in conversational recommender systems: A survey. *AI open*.

Hao, S.; Gu, Y.; Ma, H.; Hong, J. J.; et al. 2023. Reasoning with language model is planning with world model. *arXiv*.

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

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

He, T.; Liao, L.; Cao, Y.; Liu, Y.; et al. 2025a. Simulation-Free Hierarchical Latent Policy Planning for Proactive Dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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*.

Hou, Y.; Liang, X.; Zhang, J.; Yang, Q.; Yang, A.; and Wang, N. 2023. Exploring the use of invalid action masking in reinforcement learning: A comparative study of on-policy and off-policy algorithms in real-time strategy games. *Applied Sciences*.

Hu, Z.; Feng, Y.; Deng, Y.; Li, Z.; et al. 2023. Enhancing large language model induced task-oriented dialogue systems through look-forward motivated goals. *arXiv*.

Joshi, R.; Balachandran, V.; Vashishth, S.; Black, A. W.; et al. 2021. DialoGraph: Incorporating Interpretable Strategy-Graph Networks into Negotiation Dialogues. In *International Conference on Learning Representations*.

Kang, D.; Kim, S.; Kwon, T.; Moon, S.; et al. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *arXiv*.

Kang, Y.; and Cho, Y.-S. 2025. Beyond Single Emotion: Multi-label Approach to Conversational Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Korbak, T.; Perez, E.; and Buckley, C. L. 2022. RL with KL penalties is better viewed as Bayesian inference. *arXiv*.

Kwan, W.-C.; Wang, H.-R.; Wang, H.-M.; and Wong, K.-F. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*.

Levin, E.; Pieraccini, R.; and Eckert, W. 1997. Learning dialogue strategies within the markov decision process framework. In *Workshop on Automatic Speech Recognition and Understanding Proceedings*.

Li, X.; Dai, Y.; Yang, Z.; Chi, J.; et al. 2025. Utterance-level Emotion Recognition in Conversation with Conversation-level Supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liang, L.; Huang, W.; Zhang, H.; Dai, Z.; et al. 2023. Enhancement of Distribution Network Resilience: A Multi-Buffer Invalid-Action-Mask Double Q-Network Approach for Distribution Network Restoration. In *International Conference on New Energy and Power Engineering*.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; et al. 2021. Towards Emotional Support Dialog Systems. In *In Annual Meeting of the Association for Computational Linguistics.*

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*.

Luo, F.; Xu, T.; Lai, H.; Chen, X.-H.; et al. 2022. A Survey on Model-based Reinforcement Learning. *ArXiv*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; et al. 2013. Playing atari with deep reinforcement learning. *arXiv*.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; et al. 2022. Training language models to follow instructions with human feedback. *ArXiv*.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*.

Song, I.; Pendse, S. R.; Kumar, N.; and De Choudhury, M. 2024. The typing cure: Experiences with large language model chatbots for mental health support. *arXiv*.

Stasaski, K.; Kao, K.; and Hearst, M. A. 2020. CIMA: A Large Open Access Dialogue Dataset for Tutoring. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*.

Stolz, R.; Krasowski, H.; Thumm, J.; Eichelbeck, M.; et al. 2024. Excluding the irrelevant: Focusing reinforcement learning through continuous action masking. *Advances in Neural Information Processing Systems*.

Tesauro, G.; et al. 1995. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3): 58–68.

Tokic, M. 2010. Adaptive ε-greedy exploration in reinforcement learning based on value differences. In *Annual conference on artificial intelligence*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*.

Varricchione, G.; Alechina, N.; Dastani, M.; De Giacomo, G.; et al. 2024. Pure-past action masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; et al. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv*.

Wang, S.; Zhou, K.; Lai, K.; and Shen, J. 2020. Task-Completion Dialogue Policy Learning via Monte Carlo Tree Search with Dueling Network. In *Conference on Empirical Methods in Natural Language Processing*.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; et al. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *In Annual Meeting of the Association for Computational Linguistics.*

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning*.

Yan, X.; Song, Y.; Feng, X.; Yang, M.; et al. 2024. Efficient Reinforcement Learning with Large Language Model Priors. *arXiv*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*.

Ye, N.; Yang, H.; Siah, A.; and Namkoong, H. 2024. Pretraining and in-context learning IS Bayesian inference a la De Finetti. *arXiv*.

Yu, X.; Chen, M.; and Yu, Z. 2023. Prompt-Based Monte-Carlo Tree Search for Goal-Oriented Dialogue Policy Planning. In *Conference on Empirical Methods in Natural Language Processing*.

Zare, M.; Kebria, P. M.; Khosravi, A.; and Nahavandi, S. 2024. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*.

Zhang, C.; Yang, K.; Hu, S.; Wang, Z.; et al. 2023. Proagent: Building proactive cooperative ai with large language models. *arXiv*.

Zhang, Q.; Naradowsky, J.; and Miyao, Y. 2023. Ask an Expert: Leveraging Language Models to Improve Strategic Reasoning in Goal-Oriented Dialogue Models. *arXiv*.

Zhang, T.; Huang, C.; Deng, Y.; Liang, H.; et al. 2024. Strength Lies in Differences! Improving Strategy Planning for Non-collaborative Dialogues via Diversified User Simulation. *arXiv*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; et al. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv*.

Zhang, Z.; Liao, L.; Zhu, X.; Chua, T.-S.; et al. 2020. Learning Goal-oriented Dialogue Policy with opposite Agent Awareness. *ArXiv*.

Zhao, S.; Brekelmans, R.; Makhzani, A.; and Grosse, R. 2024. Probabilistic inference in language models via twisted sequential monte carlo. *arXiv*.

Zhao, W.; Zhao, Y.; Lu, X.; Wang, S.; et al. 2023. Is ChatGPT Equipped with Emotional Dialogue Capabilities? *ArXiv*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; et al. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*.

Zheng, Z.; Liao, L.; Deng, Y.; and Nie, L. 2023b. Building emotional support chatbots in the era of llms. *arXiv*.

Zhou, Y.; He, H.; Black, A. W.; and Tsvetkov, Y. 2019. A Dynamic Strategy Coach for Effective Negotiation. In *Proceedings of the SIGDialMeeting on Discourse and Dialogue*.

Zhou, Y.; Tsvetkov, Y.; Black, A. W.; and Yu, Z. 2020. Augmenting Non-Collaborative Dialog Systems with Explicit Semantic and Strategic Dialog History. In *International Conference on Learning Representations*.