Dream to Chat: Model-based Reinforcement Learning on Dialogues with User Belief Modeling

Anonymous ACL submission

Abstract

World models have been widely utilized in robotics, gaming, and auto-driving. However, their applications on natural language tasks are relatively limited. In this paper, we construct the dialogue world model, which could predict the user's emotion, sentiment, and intention, and future utterances. By defining a POMDP, we argue emotion, sentiment and intention can be modeled as the user belief and solved by maximizing the information bottleneck. By this user belief modeling, we apply the model-based reinforcement learning framework to the dialogue system, and propose a framework called DreamCUB. Experiments show that the pre-trained dialogue world model can achieve state-of-the-art performances on emotion classification and sentiment identification, while dialogue quality is also enhanced by joint training of the policy, critic and dialogue world model. Further analysis shows that this manner holds a reasonable exploration-exploitation balance and also transfers well to out-of-domain scenarios such as empathetic dialogues.

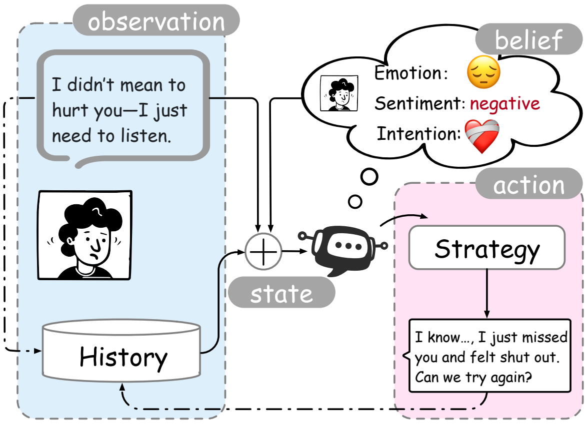

Figure 1: Paradigm of DreamCUB, in which we introduce **user belief modeling**, to speculate the unobservable state in dialogue. State becomes the union of observation and belief, which further enhances the policy.

1 Introduction

Due to strong capabilities, modern Large Language models (LLM) have obtained remarkable progress on dialogue systems (Kang et al., 2024; Zhou et al., 2024a). Among the training pipeline of conversational LLM, reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) is an important post-training stage which bootstraps the human preference and achieves a deeper alignment by interactive sampling. Although PPO (Schulman et al., 2017) is employed as the usual approach, its variants, such as DPO and GRPO, are also proposed to improve the dialogue policy. However, reinforcement learning (RL) is often subject to low sampling efficiency, high performance variance, and high computational overhead. When applied to the dialogue systems, these issues become more challenging when the model size is large and the annotation is consuming.

To alleviate these issues, Model-Based Reinforcement Learning (MBRL) (Sutton, 1991; Deisenroth and Rasmussen, 2011) is proposed, which enables the agent to learn the environment model and use it to simulate, plan, and act. Combining with recent progress on World Models (WM) (Ha and Schmidhuber, 2018), MBRL has been a power solution for visual control (Hafner et al., 2020), game (Hafner et al., 2019), auto-driving (Gao et al., 2024) and also dialogue system (Peng et al., 2018; Xu et al., 2025). For example, DDQ (Peng et al., 2018) proposes the world model of dialogue which can predict the dialogue contents. However, dialogues are highly sensitive on human psychological states, such as emotion and sentiment (Firdaus et al., 2023; Qian et al., 2023). People's reasoning, expression and intention can be affected and drifted by these inner states. However, such states are unobservable, while current MBRL studies on dialogues are based on observable states only, *i.e.*, utterances. On the other hand, previous research on empathetic dialogue systems has mostly focused on generating responses given certain emotions. However, being empathetic not only requires the ability of generating emotional responses, but more importantly, requires the understanding of user emotions and replying appropriately (Lin et al., 2019).

To bridge these gaps, in this paper, we introduce the user belief modeling into the MBRL framework, to provide a more thorough understanding of the dialogue policy. Such user beliefs may include emotion, sentiment and intention, which are unobservable states for the agent, forming a Partially Observable Markov Process (PODMP). Correspondingly, our Dialogue World Model (DWM) can not only generate future dialogue utterances, but also recognize user beliefs and behave as the reward model. To solve this problem, we refer to the theoretical derivations of POMDP-based MBRL studies (Chen et al., 2022), and deduce the DWM-RL algorithm based on the information bottleneck. Combining user belief modeling, DWM and MBRL, we propose the framework called textbfDream to **C**hat with **U**ser **B**elief (**DreamCUB**). DreamCUB simulates user belief and emotional dynamics over the course of interaction. Rather than relying on static emotion classification or purely supervised generation, DreamCUB enables an agent to imagine possible future dialogue trajectories, reason about long-term emotional impact, and plan supportive responses accordingly. Figure 1 illustrates the paradigm of DreamCUB. We summarize our contributions as follows:

- We redefine the Dialogue World Model which models user beliefs, to capture the sentimental and emotional dynamics.
- We introduce **DreamCUB**, a model-based reinforcement learning framework to apply the knowledge of Dialogue World Model on dialogue systems.
- We empirically validate our approach on daily and empathetic dialogue datasets, showing accurate emotional predictions, high response quality and strong generalizations.

2 Preliminaries

POMDP. A Partially Observable Markov Decision Process (POMDP) models the decision-making process under uncertainty when the system state is not fully observable. It is defined as 5-tuple:

$$\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R})$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $\mathcal{T}(s'|s,a)$ is the transition model, and $\mathcal{R}(s)$ is the reward function.

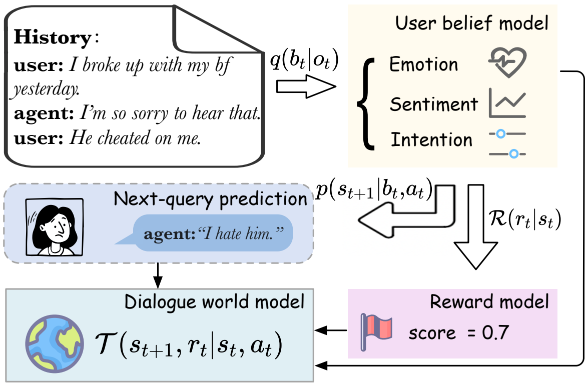

Figure 2: The dialogue world model (DWM) $\mathcal{T}(s_{t+1}, r_t|s_t, a_t)$ consists of three parts, the user belief model $q(b_t|o_t)$, the next-query prediction model $p(s_{t+1}|b_t, a_t)$ and the reward model $\mathcal{R}(r_t|s_t)$.

Reward modeling. Application of RL on textual environments requires Reward Model (RM)(Ouyang et al., 2022), which is trained from pairwise preference data (x, y_+, y_-) with x as the input, y_+ and y_- are positive and negative responses. RM is usually implemented by an LLM with the classification head added, which produces a 0-1 score. Its loss can be derived from human preference distribution by the Bradley-Terry (Bradley and Terry, 1952) model

$$\mathcal{L}_{\mathcal{R}} = \frac{1}{N}\sum_{i=1}^{N} \log \sigma(\mathcal{R}(y_+^i|x^i) - \mathcal{R}(y_-^i|x^i)) \quad (1)$$

where $\mathcal{R}$ denotes RM, $\mathcal{L}$ is the loss, and σ is the sigmoid function.

RLHF. The generative policy on language tasks solves the following problem:

$$\max_{\pi_\theta} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}, \boldsymbol{y}\sim\pi_\theta(\cdot|\boldsymbol{x})} \left[r_\phi(\boldsymbol{y}|\boldsymbol{x}) - \mathcal{L}_{KL}\right] \quad (2)$$

where $\mathcal{L}_{KL} = \beta D_{\mathrm{KL}}(\pi_\theta(\cdot|\boldsymbol{x}) \| \pi^{\mathrm{SFT}}(\cdot|\boldsymbol{x}))$ is the regularization term which prevents the RL policy deviated from SFT too much. One usual solution is to employ PPO (Schulman et al., 2017) to optimize the modified reward $r_\phi(\boldsymbol{y}|\boldsymbol{x}) - \beta\left(\log \pi_\theta(\boldsymbol{y}|\boldsymbol{x}) - \log \pi^{\mathrm{SFT}}(\boldsymbol{y}|\boldsymbol{x})\right)$.

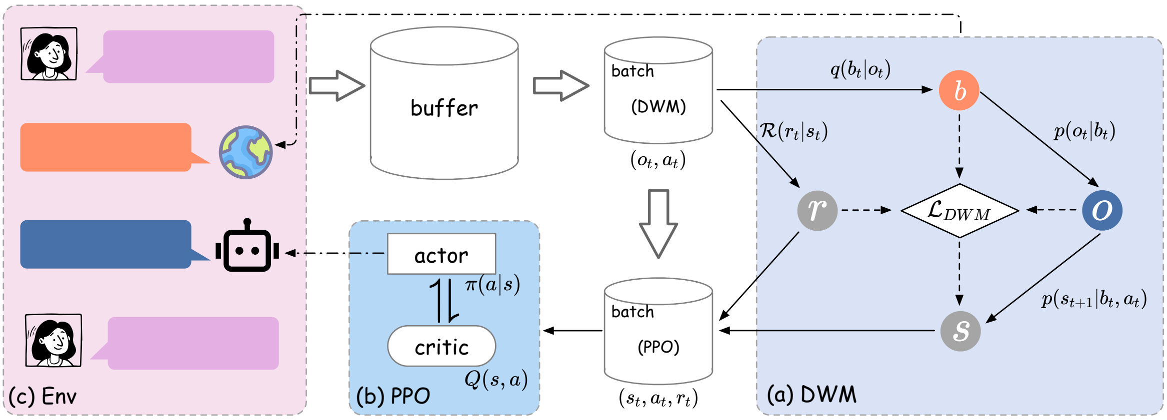

Figure 3: Training framework of DreamCUB. (a) Dynamics learning of DWM. (b) Behavior Learning of dialogue policy. (c) Interaction with environment.

3 Method

Tasks formulation. Dialogue can be characterized by an interleaved sequence of user's $query$ and agent's $response$. At the T-th turn, we denote the dialogue history as

$$hist(T) := \{query(t), resp(t)\}_{0:T-1} \quad (3)$$

where $hist$ and $resp$ abbreviate the history and response, respectively.

Recent studies usually bootstrap and annotate the agent's reply $strategy$, to have enhanced $response$ grounded by $strategy$. In this work, we further argue that the user's state, called $belief$, can also be modeling and behaving as the contextual information of subsequent $strategy$ and $response$. Such $belief$ may include the user's $emotion$, $sentiment$, and $intention$. In this formulation, the determination pipeline becomes

$$hist \oplus query \rightarrow belief \rightarrow strategy \rightarrow resp$$

System definition. The above formulation suggests $query$, $resp$, $hist$ and $strategy$ are observable to the agent while the user's $emotion$, $sentiment$ and $intention$ are unobservable. The system can then be described as a 5-tuple POMDP $(\mathcal{O}, \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T})$:

- Observation $o = (hist, query) \in \mathcal{O}$
- Belief: $b = (emotion, sentiment, intention)$
- State: $s = (o, b) \in \mathcal{S}$
- Action: $a = (strategy, resp) \in \mathcal{A}$
- Reward $r = \mathcal{R}(s)$ with s as input instead of o
- Transition Function: $\mathcal{T} := \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$.

Model implementation. To interpret this POMDP, we employ the model-based RL framework consisting of the following models:

- Belief inference model: $q(b_t|o_t)$
- Observation model: $p(o_t|b_t)$
- Belief Transition model: $p(b_{t+1}|b_t, a_t)$
- Reward model: $\mathcal{R}(r_t|s_t)$
- Actor net: $\pi(a|s)$
- Critic net: $Q(s, a)$

Taking advantage of the strong linguistic capability of LLMs, we implement all the above models based on the foundation LLM, with the prompts in three categories:

1. $q \leftarrow \text{LLM}(prompt_{cognitive})$: we implement the cognitive prompt (Wang and Zhao, 2024) for model q which allows the identification of $emotion$, $sentiment$ and $intention$.

2. $p, \pi \leftarrow \text{LLM}(prompt_{generative})$: use generative prompts for $p(o_t|b_t)$, $p(b_{t+1}|b_t, a_t)$ and the actor $\pi(a|s)$.

3. $\mathcal{R}, Q \leftarrow \text{LLM}(prompt_{classify}) \oplus \text{head}$: add the classification head on the last layer, which yields a 0-1 score (Ouyang et al., 2022).

with detailed prompt provided in Appendix A.1.

Specifically, we propose the term Dialogue World Model (**DWM**) $\mathcal{T}(s_{t+1}, r_t|s_t, a_t)$ which contains three parts: the belief inference model $q(b_t|o_t)$ which is a cognitive model to identify the user belief; the belief transition model $p(s_{t+1}|b_t, a_t) = p(b_{t+1}|b_t, a_t)p(o_t|b_t)$ which conducts the next-query generation[1], and RM $\mathcal{R}(r_t|s_t)$ which produces the reward score. These three combined together, formulating the entire DWM. Figure 2 visualizes our DWM with more details.

[1]In contrast, the dialogue policy $\pi(a|s)$ produces the next-response generation.

Algorithm 1 DWM-RL

Initialize the batch sizes B_{DWM} and B_{PPO}, the window length L and imagination horizon H
Load pretrained cognitive model q_ξ, generative model p_θ and reward model $p_\eta(r_\tau|s_\tau)$
Initialize policy $\pi_\phi(a|s)$, critic $Q_\psi(s,a)$ and the buffer $\mathcal{B} = \{\}$
while not converged **do**:
▷ *Dynamic learning*
Draw B_{DWM} data sequences $\{(o_t, a_t, r_t)\}_{t=k}^{k+L}$ from $\mathcal{B}$
Inference belief state $q_\xi(b_t|o_t)$, rollout imaginary trajectories $\{(s_\tau, a_\tau)\}_{\tau=t}^{t+H}$ with $p_\theta(s_{t+1}|b_t, a_t)$
Update ξ, θ and η by ELBO (Equation 4)
▷ *Behavior learning*
Predict rewards $p_\eta(r_\tau|s_\tau)$ for each s_τ
Draw B_{RL} data sequences $\{(s_t, a_t, r_t)\}$ from $\{(s_\tau, a_\tau, r_\tau)\}_{\tau=t}^{t+H}$
Update ϕ and ψ jointly by PPO (Equation 2)
▷ *Interact with the environment*
Get original query o_1 from dataset.
for t = $1, \ldots, T$ **do**
  Inference the belief $b_t \sim q_\xi(b_t|o_t)$, forming the state $s_t = (o_t, b_t)$
  Determine the action $a_t \sim \pi_\phi(a_t|s_t)$
  Execute a_t and get o_{t+1}, r_t
end for
Add experience to buffer $\mathcal{B} = \mathcal{B} \cup \{(s_t, a_t, r_t)\}_{t=1}^{T}$
end while

Algorithm. Posterior of beliefs and rewards, given observations and actions, can be maximized jointly by the variational information bottleneck (Tishby et al., 2000), or called the Evidence Lower Bound (ELBO) (Jordan et al., 1999):

$$\begin{aligned}&\log p(o_{1:T}, r_{1:T}|a_{1:T}) \\ &\geq \sum_{t=1}^{T} \Big(\mathbb{E}_{q(b_t|o_{\leq t}, a_{<t})}[\log p(o_t|b_t) + \log \mathcal{R}(r_t|b_t)] \\ &\quad - \mathop{\mathbb{E}}_{q(b_{t-1}|o_{t-1})}\left[D_{\mathrm{KL}}(q(b_t|o_t) \| p(b_t|b_{t-1}, a_{t-1}))\right]\Big) \doteq \mathcal{L}_{\mathrm{DWM}}\end{aligned} \tag{4}$$

with precise derivation in Appendix B.1. This lowerbound was originally proved by (Chen et al., 2022) which derives the following theorems:

Theorem 1. *The approximation error of the log-likelihood when maximizing the $\mathcal{L}_{\mathrm{DWM}}$ (the derived ELBO) defined in Equation 4 is:*

$$\begin{aligned}&\log p(o_{1:T}, r_{1:T}|a_{1:T}) - \mathcal{L}_{\mathrm{DWM}} \\ &= \mathop{\mathbb{E}}_{q(b_{1:T}|o_{1:T}, a_{1:T-1})}\left[\Sigma_{t=1}^{T} D_{\mathrm{KL}}(q(b_t|o_t) \| \bar{p}(b_t|o_t))\right]\end{aligned} \tag{5}$$

where $\bar{p}(b_t|o_t)$ denotes the true states.

Based on aforementioned consideration, we propose Algorithm 1, the Dialogue World Model-based Reinforcement Learning (DWM-RL), which contains three stages, (i) Dynamic learning, (ii) Behavior learning and (iii) Interact with the environment. Figure 3 shows the entire framework.

4 Experiment

4.1 Settings

Implementation. Llama3.1-8B-Instruct (AI@Meta, 2024) is employed as the base model. Training is conducted on OpenRLHF (Hu et al., 2024) with $L = 1024$, $H = 16$, $B_{DWM} = 256$, $B_{PPO} = 512$, $\gamma = 0.9$, $\beta = 0.01$. The learning rate is $5.0e-7$, training epoch is 1 and the replay buffer size is 24,000. RM is trained with positive response from the original dataset and negative responses from dynamic sampling.

Datasets. For DWM pertaining, we employ three types of tasks:

1. Sentiment classification: classify either Positive or Negative from the user query. We use Amazon[2], Yelp[3], and IMDB (Maas et al., 2011) as benchmarks.

2. Sentiment intensity regression: predict a 0-1 score indicating the user's sentiment polarity[4].

[2] http://jmcauley.ucsd.edu/data/amazon/
[3] https://www.yelp.com/dataset/download
[4] 0 means fully negative and 1 means fully positive.

task →	sentiment classification						intensity regression		emotion classification			
	Amazon		IMDb		Yelp		V-reg	SST	GoEmotion		E-c	
model ↓	ACC	MaF1	ACC	MaF1	ACC	MaF1	pcc	pcc	ACC	MaF1	MiF1	MaF1
llama2-7b-chat	64.19	69.17	83.23	86.36	87.69	89.48	9.12	72.83	35.71	27.15	41.40	28.60
Emollama-chat-7b	56.95	63.43	73.52	82.90	74.46	81.01	**88.00**	82.00	37.00	**39.00**	**69.30**	**54.00**
DWM	**74.13**	**73.89**	**96.38**	**96.38**	**97.42**	**97.31**	86.38	**90.28**	**39.44**	30.41	51.32	48.67
llama2-13b-chat	69.54	71.93	90.66	91.51	90.07	91.06	24.06	81.10	27.80	33.70	42.40	30.20
Emollama-chat-13b	65.01	69.61	55.70	69.51	51.28	59.86	**88.40**	81.60	35.00	**37.00**	**69.60**	54.50
DWM	**73.84**	**73.68**	**96.69**	**96.69**	**97.53**	**97.41**	88.36	**90.66**	**37.21**	33.81	69.41	**57.73**
llama3-8b-instruct	72.38	73.92	92.63	92.66	93.21	92.94	57.04	82.17	32.83	**34.43**	43.95	41.38
DWM ($q(b\|o)$)	**87.87**	**87.87**	**96.99**	**96.99**	**96.34**	**96.17**	**86.50**	**90.19**	**33.60**	32.52	**58.39**	**59.42**

Table 1: Performance of dialogue world model compared with state-of-the-art emotional cognition models. V-reg and E-c are two subtasks of SemEval 2018 Task1. pcc denotes the Pearson correlation coefficient.

history	*user:*	*Did you hear about the robbery?*
	agent:	*Did I hear about it? I saw it happen.*
	user:	*Are you serious?*
belief	**Emotion: *"surprise"*, Sentiment:*"negative"*, *"0.388"***	
	Ground Truth	surprise, negative
	agent:	*<inform> I was there.*
query	*user:*	Predicted: **What went down?**
		Ground Truth: What happened ?

Table 2: Case of DWM on user belief cognition ($q(b_t|o_t)$) and next-query prediction ($p(o_t|b_t, o_{t-1})$). Contents from the original dataset are *italic*, and results of DWM are **bolded**.

We use Stanford Sentiment Treebank (SST) (Socher et al., 2013) and the corresponding subtask in SemEval-2018 Task1: Affect in Tweet (Mohammad and Kiritchenko, 2018).

3. Emotion classification: select the appropriate emotion from the candidates, such as joy, anger, sad, etc. We use GoEmotion (Demszky et al., 2020) and again the corresponding subtask in SemEval-2018 (Mohammad and Kiritchenko, 2018).

For PPO training, we use DailyDialogue (Li et al., 2017), ESconv (Liu et al., 2021), EmpatheticDialogues (Rashkin et al., 2019). The first two have annotations of emotion, strategy and response, while the last one only has annotations of emotion and response. To gain significant generalizability, we use DailyDialogue (Li et al., 2017), which is focused on daily topics, as both training and in-domain (ID) test sets. The other two, which are more focused on empathetic dialogue, are used for out-of-domain (OOD) evaluation purposes only.

Metrics. For classification tasks, we employ the metrics of accuracy (ACC), Micro-F1 (MiF1) and Macro-F1 (MaF1). We also refer the evaluation methods proposed by Kang et al. (2024), which propose the $bias$ based on Bradley-Terry model (Bradley and Terry, 1952). Smaller $bias$ means less bias, therefore is better. For regression tasks, we use the Pearson correlation coefficient (pcc). For generation task, we utilize the famous Bleu-2 (B-2), Rouge-L (R-L) and Distinct-2 (D-2). The first two are similarity-based metrics, while the last one encourages the response diversity. We also conduct human annotations to evaluate the responses. We leave the annotation principle, and metric details in the Appendix.

4.2 Training of DreamCUB

Figure 4 visualizes the training curves, which shows that our Algorithm 1 converges and the return can be maximized. More specifically, Figure 4 (bottom-right) highlights a preference evolution of the dialogue policy, the response length. At the beginning of training, the LLM tends to provide long responses, which are not natural enough considering the daily conversation situation. As joint training with DWM, the responses start to become shorter, and finally reaching a balance.

4.3 Results of dialogue world model

Emotion Cognition. Table 1 shows our DWM after the pretraining. We achieve state-of-the-art accuracy on all three types of emotional cognitive tasks, surpassing the base model and EmoLLama. To be consistent with our RL training, we use the Llama3-based version for the subsequent formal experiments. Table 2 shows a good case of emotion cognition.

Dialogue Generation. Our system transition model (p) of DWM needs to predict the user intention or query, based on the current conversation

Method	Emotion			Strategy			Response		
	ACC	MaF1	*bias* ↓	ACC	MaF1	*bias* ↓	B-2	R-L	D-2
Direct	-	-	-	52.60	18.03	1.66	3.35	10.33	44.74
+ Retrieve	-	-	-	30.92	21.17	0.67	2.78	9.67	40.60
+ Refine	-	-	-	48.27	28.28	0.70	2.56	8.70	43.67
+ Self-Refine	-	-	-	49.76	22.15	1.18	2.40	7.75	34.01
+ CoT	-	-	-	38.94	29.99	**0.27**	1.78	6.00	**55.26**
+ FSM	73.01	24.50	1.63	46.86	21.22	1.30	2.70	9.44	38.75
+ SFT	76.76	14.35	2.03	60.19	44.82	0.82	6.81	18.52	43.36
+ CoT + SFT	83.48	15.60	1.98	60.11	44.90	0.66	6.61	18.07	42.87
+ FSM + SFT	83.28	14.44	2.22	64.05	48.36	0.62	5.85	21.77	47.43
+ DreamCUB (ours)	**88.05**	**50.88**	**0.74**	**67.80**	**62.29**	0.33	**11.65**	**29.09**	49.36

Table 3: ID results on automatic metrics on DailyDialogue, including classification metrics such as Accuracy (ACC), Macro-F1 (MaF1) and $bias$, and generation metrics such as BLEU-2 (B-2), ROUGE-L (R-L) and Distinct-2 (D-2). The best results of each LLM are **bolded** and the second best are underlined.

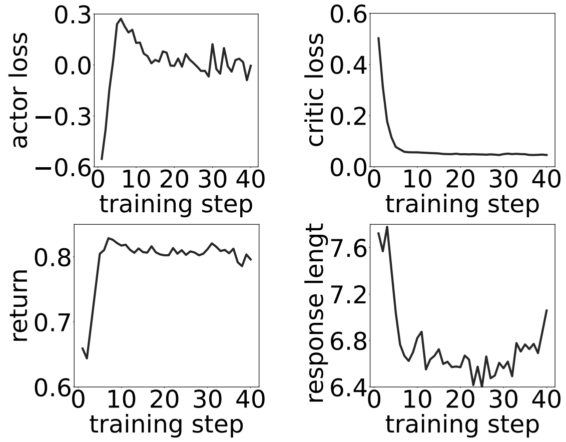

Figure 4: Training plots of DreamCUB, including the actor loss (top-left), the critic loss (top-right), return (bottom-left) and reward (bottom-right).

context. However, next-query prediction is difficult to have qualitative results, since user queries could be open topics. Instead, Table 2 shows a typical case of p. One can observe that p can understand contextual information, and generate reasonable user queries which sometimes are similar to the ground truth.

Scalability. Table 1 also shows results of the 13B-based experiment, in which our DWM still perform better than the base model and EmoLlama on most of the metrics, suggesting our method are scalable to higher model and data sizes.

4.4 Results of Dialogue Policy

Baselines. We consider the following baselines:
(1) Direct: directly inference the LLM, with the same context.
(2) Retrieve: use RAG (Fan et al., 2024) to retrieve the top-2 strategy. We employ E5-large (Wang et al., 2024b) as the semantic retriever.
(3) Refine: a straightforward refinement method in which the model revises its initial response to incorporate emotional support considerations.
(4) Self-Refine: a method (Madaan et al., 2023) initiates by generating feedback emphasizing emotional support from the initial response, then refining the response based on this feedback.
(5) CoT: uses the Chain-To-Thought prompt (Wei et al., 2022), which first generate the seeker's *emotion*, which then guides the generation of strategy and response.
(6) FSM: the finite state machine (Wang et al., 2024c) with finite sets of states and state transitions triggered by inputs, and associated discrete actions.

Results. Table 3 shows the ID results of our dialogue policy $\pi(o)$, on the classification of emotion and strategy, as well as metrics of response. For most prompt-based baselines, it is difficult to classify the user emotion without pretrained knowledge, therefore we do not list this part of results. The only exception is FSM, which provides a detailed, situational strategy for the model to inference the emotion and strategy from finite sets. On the other hand, the finetuning-based baselines can classify both user emotion and the assistant strategy, with the training datasets organized accordingly. Nevertheless, our DreamCUB consistently outperforms these baselines, on both emotion, strategy and response. Note we consider both similarity-based metrics (B-2 and R-L) and diversity-based metrics (D-2) here, which indicates a reasonable balance achieved by DreamCUB. Table 11 and 12 in the Appendix further shows per-emotion and per-strategy results, indicating DreamCUB behaves

	Method	Emotion			Strategy			Response		
		ACC	MaF1	$bias\downarrow$	ACC	MaF1	$bias\downarrow$	B-2	R-L	D-2
ESconv	SFT	25.12	11.38	2.65	11.15	5.54	2.19	3.30	12.90	27.67
	CoT + SFT	32.90	**15.48**	2.21	15.28	8.09	**1.75**	2.33	9.00	31.13
	FSM + SFT	30.23	6.84	2.62	18.76	8.12	1.88	2.70	10.46	28.10
	DreamCUB (ours)	**34.26**	14.78	**1.94**	**30.78**	**10.90**	1.80	**3.68**	**13.71**	**33.23**
Empathetic -Dialogues	SFT	4.03	1.44	5.44	N/A	N/A	N/A	2.56	7.68	34.83
	CoT + SFT	12.20	7.77	**3.60**	N/A	N/A	N/A	2.56	9.81	**39.39**
	FSM + SFT	4.59	2.20	5.57	N/A	N/A	N/A	2.61	9.87	30.52
	DreamCUB (ours)	**16.49**	**17.58**	5.15	N/A	N/A	N/A	**4.03**	**13.15**	37.08

Table 4: OOD results on automatic metrics on ESconv and EmpatheticDialogues, including classification metrics such as Accuracy (ACC), Macro-F1 (MaF1) and $bias$, and generation metrics such as BLEU-2 (B-2), ROUGE-L (R-L) and Distinct-2 (D-2). The best results of each LLMs are **bolded** and the second best are underlined.

Method	Fluency	Emotion	Acceptance	Effectiveness	Sensitivity	Alignment	Satisfaction
Llama3-8B-Instruct	2.95	3.00	2.60	2.40	2.70	2.70	2.60
+ Refine	3.09	3.09	2.73	2.91	2.91	2.82	2.84
+ Self-Refine	3.10	3.15	2.80	2.70	2.90	2.80	2.80
+ CoT	3.08	3.08	2.83	2.67	3.00	2.83	2.83
+ FSM	3.30	3.35	2.90	2.90	3.00	2.90	2.93
+ SFT	3.15	3.40	2.70	2.70	2.90	3.30	2.90
+ CoT + SFT	3.67	**3.61**	3.22	3.67	3.56	3.35	3.45
+ FSM + SFT	3.80	3.55	3.40	3.70	3.80	3.70	3.65
+ DreamCUB	**3.85**	3.52	**4.09**	**3.90**	**3.86**	**4.01**	**3.98**

Table 5: Human evaluation of response quality on ESconv and EmpatheticDialogues.

equally across different emotions and strategies.

Table 4 further shows the OOD results on esconv and empathetic dialogues, from models trained by DailyDialogue, conversations of daily topics. In this situation, DreamCUB still generally performs better than baselines, with seldom exceptions. This observation ensures that the knowledge learned from general dialogues can smoothly transfer to some specific domains, *i.e.*, the emphatic dialogues, with the assistance of the dialogue world model, as well as model-based reinforcement learning.

Human evaluations. Table 5 presents human evaluation results. DreamCUB achieves the highest scores in almost all aspects, which verifies the automatic evaluation results.

Cases. Table 6 showcases the dialogue policy and the dialogue world model, after being jointly trained by Algorithm 1. From this typical case, it can be observed that both π and q adapt to the specific domain well, provide accurate emotion and sentiment identification, and finally generate reasonable strategies and responses.

4.5 Ablation

To verify the effectiveness of the components of DreamCUB, here we consider the following ablation settings:

- w/o WB: train the PPO policy without the knowledge of DWM.
- w/o RL: inference DWM directly, without the RL training.
- w/o b in DWM: do not consider the user belief in the dialogue world model, *i.e.*, only use dialogue history to predict the next-query of user.
- w/o b in $\mathcal{R}$: do not consider the user belief in the reward model, *i.e.*, provide the reward score based on the dialogue context only.
- singe-model: make the policy and DWM a single parameter-shared model.

As shown in Table 7, DreamCUB still performs the best on all the metrics, suggesting all its components are necessary to reach optimal performance. Especially, DreamCUB utilizes the user belief information (b), resulting in further performance benefit compared to w/o b in DWM and $\mathcal{R}$. Nevertheless, both w/o b in DWM and w/o b in $\mathcal{R}$ can still surpass the baselines in Table 3, indicating the pure application of MBRL on dialogue system can substantially improve the performance. Last, the single-model approach can not behave as good as DreamCUB, which indicates that it is still better to use separate models for the dialogue policy and the world model, given the current setting.

Characters	Emotion	Sentiment	Strategy	Response
User	*no emotion*	-	-	*Hi there, I want to reserve a hotel room.*
Assistant	-	-	*question*	*No problem at all. Could I have your full name, please?*
User	*no emotion*	-	-	*Sure, John Sandals.*
Assistant	-	-	*question*	*Hi, Mr.Sandals. I'm Michelle, at your service. When do you need the room?*
User	*no emotion*	-	-	*My plans are to be there April 14th to the 17th.*
Assistant	-	-	*question*	*We have new room rates, sir. Will that be acceptable to you?*
User	*surprise*	-	-	*Well, it depends on the price, of course. What is it ?*
Assistant	-	-	*inform*	*It's $ 308 a night.*
Direct	no emotion	0.5	-	Our nightly rate currently sits at $375. If you'd like, I can provide details about available amenities as well.
DreamCUB	**surprise**	**0.43**	**inform**	**The price is $ 458 per night.**
User	*no emotion*	-	-	*I have no problem with that.*
Assistant	-	-	*question*	*Great ! Would you prefer smoking or nonsmoking?*
Direct	no emotion	0.5	-	Perfect, I'll make a note of that. By the way, do you have any special requests or preferences during your stay?
DreamCUB	**happiness**	**0.78**	**question**	**Would you like a smoking or non-smoking room?**
User	*disgust*	-	-	*Definitely nonsmoking . I can't handle that smell.*
Assistant	-	-	*question*	*Nonsmoking. Now, is a queen-size bed okay?*
Direct	digust	0.6	-	Thanks for letting me know. Before we proceed, could you please confirm your email address?
DreamCUB	**disgust**	**0.19**	**directive**	**Nonsmoking, nonsmoking. Now, if you'll just give me your phone number.**

Table 6: Typical cases generated by DreamCUB in Dailydialogue. Contents from the original dataset are *italic*, and results of DreamCUB are **bolded**.

Method	Emotion			Strategy			Response		
	ACC	MaF1	*bias* ↓	ACC	MaF1	*bias* ↓	B-2	R-L	D-2
w/o WB	87.67	43.36	0.94	62.13	53.53	0.79	4.96	17.93	42.57
w/o RL	80.31	23.75	0.78	63.61	56.87	0.51	5.13	18.27	42.54
w/o b in p	86.71	41.36	1.19	61.13	52.68	0.54	6.16	19.26	42.75
w/o b in $\mathcal{R}$	87.86	48.43	0.94	64.09	55.19	1.03	11.04	28.64	**49.55**
single-model	86.79	38.03	1.45	58.26	45.02	0.86	4.87	17.74	41.04
DreamCUB (ours)	**88.05**	**50.88**	**0.74**	**67.80**	**62.29**	**0.33**	**11.65**	**29.09**	49.36

Table 7: Ablation study on DailyDialogue. The best results of each LLMs are **bolded** and the second best are underlined.

5 Related Work

RL on dialogue system. RL enhances dialogue systems in instruction following, task completion, reasoning, and emotional expression. Methods like RLHF (Ouyang et al., 2022) align models with human feedback via PPO, while Q-star (Wang et al., 2024a) improves reasoning through multi-step Q-learning. DQ-HGAN (Li et al., 2024) uses graph attention for emotionally supportive responses, and ArCHer (Zhou et al., 2024b) applies hierarchical RL for better multi-turn planning. In our method, we leverage a world model to enrich the inference of emotional and situational states.

World Models. World Models (Ha and Schmidhuber, 2018) focus on high-dimensional inputs, with PlaNet (Hafner et al., 2019) and Dreamer (Hafner et al., 2020) using latent rollouts for efficient decision-making. MBRL focuses on building world models for planning, policy optimization, and uncertainty-aware control. Offline methods such as MOPO (Yu et al., 2020) and MOReL (Kidambi et al., 2021) add uncertainty constraints for safety. Our method models emotion and context as latent variables, using a world model to enhance dialogue state transitions.

6 Conclusion

In this paper, we propose a framework called DreamCUB, to introduce the MBRL on the dialogue system, with user belief modeling of emotion, sentiment and intention. We first pretrain a dialogue world model which allows the user emotional identification and the next-query prediction, then jointly train this world model with dialogue policy, to achieve better performance on the daily dialogues. We further verify the effectiveness of user belief both in the world model and the reward model, as well as the typical conversation cases.

7 Limitation

Due to time and page limits, here we only explore a limited subset of user beliefs, including emotion, sentiment, and intention. Nevertheless, user belief modeling has the potential to consider more features, for example, user preference, habit, and memory. A more thorough user modeling might further enhance the performance.

In addition to dialogue, language tasks have versatile scenarios, including question-answering, translation, summarization, and textual games. We expect this study could be a starting point of the world model application on textual environments, which may step ahead on generalist artificial intelligence.

8 Ethical Considerations

DreamCUB models the user beliefs, which might be correlated with the user's private information. Therefore, the confidentiality of datasets needs to be strictly confirmed. Also, DreamCUB can exhibit the user beliefs on the screen, which also has the potential of user inconvenience. Users should be warned of this condition before using industrial applications.

References

AI@Meta. 2024. Llama 3 model card.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Xiaoyu Chen, Yao Mark Mu, Ping Luo, Shengbo Li, and Jianyu Chen. 2022. Flow-based recurrent belief state learning for POMDPs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3444–3468. PMLR.

Marc Peter Deisenroth and Carl Edward Rasmussen. 2011. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 465–472, Madison, WI, USA. Omnipress.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Mauzama Firdaus, Gopendra Singh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Multi-step prompting for few-shot emotion-grounded conversations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 3886–3891, New York, NY, USA. Association for Computing Machinery.

Yinfeng Gao, Qichao Zhang, Da-Wei Ding, and Dongbin Zhao. 2024. Dream to drive with predictive individual world model. *IEEE Transactions on Intelligent Vehicles*, pages 1–16.

David Ha and Jürgen Schmidhuber. 2018. World Models.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2020. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2019. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565. PMLR.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2021. MOReL : Model-Based Offline Reinforcement Learning. *Preprint*, arXiv:2005.05951.

Ge Li, Mingyao Wu, Chensheng Wang, and Zhuo Liu. 2024. DQ-HGAN: A heterogeneous graph attention network based deep Q-learning for emotional support conversation generation. *Knowledge-Based Systems*, 283:111201.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

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

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

M. E. J. Newman. 2023. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

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

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

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2192, Melbourne, Australia. Association for Computational Linguistics.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think twice: A human-like two-stage conversational agent for emotional response generation. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 727–736, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Richard S. Sutton. 1991. Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bull.*, 2(4):160–163.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024a. Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning. *Preprint*, arXiv:2406.14283.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Xiaochen Wang, Junqing He, Zhe yang, Yiru Wang, Xiangdi Meng, Kunhao Pan, and Zhifang Sui. 2024c. FSM: A Finite State Machine Based Zero-Shot Prompting Paradigm for Multi-Hop Question Answering. *Preprint*, arXiv:2407.02964.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Kai Xu, Zhenyu Wang, Yangyang Zhao, and Bopeng Fang. 2025. An efficient dialogue policy agent with model-based causal reinforcement learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7331–7343, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. 2020. MOPO: Model-based Offline Policy Optimization. *Preprint*, arXiv:2005.13239.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024a. Think before you speak: Cultivating communication skills of large language models via inner monologue. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3925–3951, Mexico City, Mexico. Association for Computational Linguistics.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024b. ArCHer: Training Language Model Agents via Hierarchical Multi-Turn RL. *Preprint*, arXiv:2402.19446.

A More Implementation Details

A.1 Prompts

Prompt of DWM ($q(b_t|o_t)$). The following prompt is utilized by the DWM model for emotion inference tasks.

$prompt_{cognitive}$:

Below is a dialogue between a user and an assistant. The dialogue history is enclosed within <history> tags.
<history> {history} </history>
The user's current emotion before the assistant's last reply is: {emotion}.
The assistant's reply, employing the {strategy} strategy, is: {assistant reply}
Your task is to analyze the user's mental belief **after** receiving the assistant's reply. Complete the following three tasks based on the updated user emotion:
1. **Sentiment classification:** Classify the user's emotional polarity as either: -1 = negative, 0 = neutral, 1 = positive. Output format: {"sentiment_class": int}
2. **Sentiment intensity regression:** Estimate the user's overall sentiment as a real number between 0 (extremely negative) and 1 (extremely positive). Output format: {"sentiment_score": float}
3. **Emotion classification:** Classify the user's emotion into one or more of the following categories: {no emotion, happiness, surprise, fear, disgust, sadness, anger}. Output format: {"emotions": ["emotion1", "emotion2", ...]}

Prompt of DWM ($p(s_{t+1}|b_t, o_t)$). The following prompt is utilized by the DWM model for next-query prediction.

$prompt_{generative}$:

Below is a dialogue between a user and an assistant. The dialogue history is enclosed within <history> tags.
<history>
{history}
</history>
The user's current emotion before the assistant's last reply is: {emotion}.
The assistant's reply, employing the {strategy} strategy, is:
{assistant reply}
If you are the user:
1. Give the user's response after receiving this reply:
{user response}

Based on the updated user emotion after receiving the assistant's reply, complete the following tasks:
2. Sentiment classification:
Classify the user's emotional polarity as either:
-1 = negative, 0 = neutral, 1 = positive
Output format: {"sentiment_class": int}
3. Sentiment intensity regression:
Estimate the user's overall sentiment as a real number between 0 (extremely negative) and 1 (extremely positive).
Output format: {"sentiment_score": float}
4. Emotion classification:
Classify the user's emotion into one or more of the following categories: {no emotion, happiness, surprise, fear, disgust, sadness, anger} Output format: {"emotions": ["emotion1", "emotion2", ...]}

$prompt_{RL}$:

Below is a dialogue between a user and an assistant. The dialogue history is enclosed within <history> tags.
<history> {history} </history>
User's emotion: {belief}
Given the user's emotion and the dialogue so far, first infer the most appropriate assistant strategy to move the dialogue forward.
Then, using the inferred strategy, the user's emotion, and the dialogue history, generate the next assistant response that naturally continues the dialogue.
Please output in the following format:
Assistant's strategy: {strategy}
Assistant's response: {response}

Prompt of Actor, Critic and RM. This prompt guides the assistant to first infer an appropriate conversational strategy based on the user's emotional state and dialogue history, and then generate a fitting response that aligns with that strategy.
The Critic and Reward model's prompt should be aligned with the Actor's in order to accurately evaluate the state value and reward.

A.2 Details of Datasets

Table 8 presents a comparison of three widely used emotion-centric dialogue datasets: ESConv, DailyDialog, and EmpatheticDialogues. Each dataset is annotated with both emotional categories and communication strategies (where available). ESConv includes a rich set of eight emotions and a diverse set of support strategies, which are abbreviated in the table for brevity. DailyDialog provides a smaller set of emotions along with basic dialogue act types. EmpatheticDialogues focuses primarily on emotional labels, covering a broader spectrum of feelings, with only the top 10 most frequent emotions shown here. This comparison highlights the varying granularity and scope of annotations across datasets used in empathetic and emotional dialogue research.

Table 9 shows an example dialogue snippet from the ESConv dataset. It illustrates a conversation where the seeker expresses anxiety about quitting a disliked job without a secure alternative. The dialogue is annotated with the topic, the seeker's query, the emotional state (anxiety with high intensity), and the empathetic strategy used by the supporter—in this case, a "reflection of feelings." This example highlights how ESConv captures nuanced emotional expression alongside supportive conversational strategies.

Table 10 presents a comparison of key statistics across three dialogue datasets: ESConv, DailyDialogue, and EmpatheticDialogues. It includes data on the number of sessions, utterances, average utterance lengths, and speaker-specific information such as utterance counts, average lengths, and the number of annotated strategies and emotions.

A.3 Metrics of Classification and Regression

F1-scores. F1-related scores include Micro-F1 and Macro-F1. Micro-F1 considers the overall precision and recall of all instances, while Macro-F1

Dataset	Annotations	Types
ESconv	Emotion Strategy	anger, anxiety, depression, disgust, fear, nervousness, sadness, shame Que., Paraphrasing &Res., Ref., Self-Dis., Aff.& Rea., Pro., Inf., Others
DailyDialgoue	Emotion Strategy	anger, disgust, fear, happiness, sadness, surprise, no emotion inform, question, directive, and commissive
EmpatheticDialogues	Emotion	surprised, grateful, proud, sentimental, excited, sad, disgusted, angry, joyful, . . .

Table 8: Lists of emotions and strategies of ESConv, DailyDialgoue and EmpatheticDialogues. Strategies of ESconv here are abbreviated names; for full names, refer to the Appendix. Only the most frequent 10 emotions of EmpatheticDialogues are listed.

Topic	I hate my job but I am scared to quit and seek a new career.
Query	*{history}* *seeker:* Seriously! What I'm scare of now is how to secure another job.
Emotion	Anxiety (intensity: 5)
Strategy	Reflection of feelings
Response	*supporter:* I can feel your pain just by chatting with you.

Table 9: An example of *ESconv*.

Category		ESconv	DailyDialogue	EmpatheticDialogues(test)
# Sessions		1.3K	13.1k	2.5K
# Utterances		38K	103.0k	11.0K
Average # Utterances		28.9	7.9	4.3
Average Utterance Length		18.8	13.6	16.7
Seeker/Speaker1	# Utterances	20K	53.8k	5.7K
	Avg # Utterances	15.4	4.1	2.2
	Avg Uttr Len	16.8	13.2	20.8
	# Strategies	-	4	-
	# Emotions	11	7	32
Supporter/Speaker2	# Utterances	18K	49.2k	5.2K
	Avg # Utterances	13.6	3.9	2.1
	Avg Uttr Len	21.0	14.1	12.3
	# Strategies	8	4	-
	# Emotions	-	7	32

Table 10: Statistics of ESConv, DailyDialogue, EmpatheticDialogues.

equals the average F1-score of labels.

bias. We define the preference *bias* as *how much the model prefers certain labels over others*. To quantify the preference for each strategy in LLMs, we employ the Bradley-Terry model (Bradley and Terry, 1952), which is widely used in human preferencemodeling (Rafailov et al., 2023). Following Newman (2023), we formally derive the preference p for strategy i as follows:

$$p_i' = \frac{\sum_j (w_{ij} p_j)/(p_i + p_j)}{\sum_j w_{ji}/(p_i + p_j)} \tag{6}$$

where w_{ij} represents the number of times the model predicts strategy i when the ground-truth strategy is j. All of the preference p_i are initialized as 1 and updated through iteration of the Eq (6) , where p_i' represents the preference in the next iteration. After the final iteration, we scale the total sum of p_i to 8 ($\sum p_i = 8$) so that the average $\bar{p}$ becomes 1, indicating a strong preference for strategy i if $p_i > 1$.

We use a standard deviation of preferences p_i across the strategies as $bias$.

$$bias = \sqrt{\frac{\sum_{i=1}^{N} (p_i - \bar{p})^2}{N}} \tag{7}$$

where a higher value for $bias$ indicates that the model exhibits a clear preference for both preferred and non-preferred strategies (Kang et al., 2024).

Pearson Correlation Coefficient. The Pearson correlation coefficient r provides a dimensionless index of the linear relationship between two continuous variables x and y. Formally, r is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{8}$$

A.4 Metrics of Generation

Bleu-2. B-2(Papineni et al., 2002) first compute the geometric average of the modified n-gram precisions, p_n, using n-grams up to length N and positive weights w_n summing to one.

Next, let c be the length of the prediction and r be the reference length. The BP and Bleu-2 are computed as follows.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}. \tag{9}$$

$$\text{Bleu} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \tag{10}$$

Rouge-L. R-L(Lin, 2004) propose using LCS-based F-measure to estimate the similarity between two summaries X of length m and Y of length n, assuming X is a reference summary sentence and Y is a candidate summary sentence, as follows:

$$\begin{aligned} R_{lcs} &= \frac{LCS(X,Y)}{m} \\ P_{lcs} &= \frac{LCS(X,Y)}{n} \\ F_{lcs} &= \frac{(1+\beta^2)\, R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \end{aligned} \tag{11}$$

Where $\text{LCS}(X,Y)$ is the length of a longest common subsequence of X and Y, and $\beta = P_{lcs}/R_{\text{lcs}}$ when $\partial F_{lcs}/\partial R_{lcs} = \partial F_{lcs}/\partial P_{lcs}$. In DUC, β is set to a very big number $(\rightarrow \infty)$. Therefore, the LCS-based F-measure, i.e. Equation 11, is Rouge-L.

Dist-2. Li et al. (2015) report the degree of diversity by calculating the number of distinct unigrams and bigrams in generated responses. The value is scaled by the total number of generated tokens to avoid favoring long sentences:

$$Dist(n) = \frac{Count(unique\ n-gram)}{Count(n-gram)} \tag{12}$$

A.5 Principle of Human Scoring

We start with the criteria proposed by Kang et al. (2024). The human evaluation is aimed to algin with the ultimate purpose of ESC, the seeker's *satisfaction*. To achieve this, the supporter's behavior can be further classified into the following criteria:
Acceptance: Does the seeker accept without discomfort;
Effectiveness: Is it helpful in shifting negative emotions or attitudes towards a positive direction;
Sensitivity: Does it take into consideration the general state of the seeker. Furthermore, to clarify the capability of LLMs to align strategy and responses, we include Alignment.

To achieve a more elaborate assessment, we consider three more dimensions addressing the generation quality:
Fluency: the level of fluency of response.
Emotion: the emotional intensity of response which could affect the seeker's emotion state.
Interesting: Whether the response can arouse the seeker's interest and curiosity, presenting unique ideas, vivid expressions or engaging elements that capture the seeker's attention and make the interaction more appealing.

We engage our interns as human evaluators to rate the models according to these multiple aspects, namely Fluency, Emotion, Interesting, and Satisfaction, with Satisfaction covering Acceptance, Effective, Sensitivity, and Satisfaction itself.
Throughout this evaluation process, we strictly comply with international regulations and ethical norms, ensuring that all practices conform to the necessary guidelines regarding participant involvement and data integrity.
Evaluators are required to independently evaluate each sample in strict accordance with the pre - established criteria. By adhering to these principles, the evaluation process maintains objectivity, standardization, and consistency, thus enhancing the overall quality and credibility of the evaluation results.
The detailed manual scoring criteria are as follows:

- Fluency:

  1: The sentence is highly incoherent, making it extremely difficult to understand and failing to convey a meaningful idea.

  2: The sentence has significant incoherence issues, with only parts of it making sense and struggling to form a complete thought.

  3: The sentence contains some incoherence and occasional errors, but can still convey the general meaning to a certain extent.

  4: The sentence is mostly fluent with only minor errors or slight awkwardness in expression, and effectively communicates the intended meaning.

  5: Perfect. The sentence is completely fluent, free of any errors in grammar, punctuation, or expression, and clearly conveys the idea.

- Emotion:

  1: The emotional expression is extremely inappropriate and chaotic, not in line with the content, and may convey wrong emotions.

  2: The emotional expression has obvious flaws, either too weak or exaggerated, and is disjointed from the content.

  3: The emotional expression is average. It can convey basic emotions but lacks depth and has minor issues.

  4: The emotional expression is good. It can effectively convey the intended emotion with an appropriate intensity and is well integrated with the content.

  5: The emotional expression is excellent. It is rich, nuanced, and perfectly matches the content, capable of evoking a strong and appropriate emotional response.

- Acceptance:

  1: The response inescapably triggers emotional resistance.

  2: The response is highly likely to trigger emotional resistance.

  3: The response has a possibility of emotional resistance occurring.

  4: The response rarely provokes emotional resistance.

  5: The response has no occurrence of emotional resistance.

- Effectiveness:

  1: The response actually worsens the seeker's emotional distress.

  2: The response carries the risk of increasing stress levels, and this outcome varies depending on the individual user.

  3: The response fails to alter the seeker's current emotional intensity and keeps it at the same level.

  4: The response shows promise in calming the emotional intensity; however, it is overly complicated or ambiguous for the user to fully comprehend and utilize effectively.

  5: The response appears to be highly effective in soothing the seeker's emotions and offers valuable and practical emotional support.

- Sensitivity:

  1: The response renders inaccurate evaluations regarding the seeker's state.

  2: The response is characterized by rash judgments, as it lacks adequate assessment and in-depth exploration of the seeker's state.

  3: The response is formulated with a one-sided judgment and a limited exploration of the seeker's state.

  4: The response demonstrates an understanding that only covers a part of the seeker's state.

  5: The response precisely grasps the seeker's state and is appropriately tailored according to the seeker's actual situation.

- Alignment:

  1: The response is in total contradiction to the predicted strategy.

  2: The response has a minor deviation from the predicted strategy.

  3: There is some ambiguity between the response and the predicted strategy.

  4: The response largely matches the predicted strategy, yet it contains some ambiguous elements.

  5: The response effectively makes itself consistent with the predicted strategy.

- Satisfaction:

1: The response is extremely disappointing. It doesn't answer the question at all and is of no help.

2: The response is poor. It only gives a partial answer and leaves many doubts unresolved.

3: The response is average. It meets the basic requirements but isn't particularly outstanding.

4: The response is good. It answers the question clearly and provides some useful details.

5: The response is excellent. It not only answers the question perfectly but also offers valuable additional insights.

B More Results

B.1 Evidence Lower Bound Derivations

The variational bound for latent dynamics models $p\left(o_{1:T}, b_{1:T} \mid a_{1:T}\right) = \prod_t p(b_t|b_{t-1}, a_{t-1})p(o_t|b_t)$ and a variational posterior $q\left(b_{1:T} \mid o_{1:T}, a_{1:T}\right) = \prod_t q\left(b_t \mid o_{\leq t}, a_{<t}\right)$ follows from importance weighting and Jensen's inequality as shown,

$$
\begin{aligned}
&\log p\left(o_{1:T}, r_{1:T} | a_{1:T}\right) \\
&= \log \mathrm{E}_{p(b_{1:T}|a_{1:T})}\left[\prod_{t=1}^{T} p\left(o_t|b_t\right) \mathcal{R}\left(r_t|b_t\right)\right] \\
&= \log \mathrm{E}_{q(\mathbf{b}|\mathbf{o},\mathbf{a})}\left[\prod_{t=1}^{T} \frac{p\left(o_t|b_t\right) p\left(b_t|b_{t-1}, a_{t-1}\right)}{q\left(b_t|o_{\leq t}, a_{<t}\right)} \mathcal{R}\left(r_t|b_t\right)\right] \\
&\geq \mathrm{E}_{q(b_{1:T}|o_{1:T}, a_{1:T})}\left[\sum_{t=1}^{T} \log p\left(b_t|b_{t-1}, a_{t-1}\right)\right. \\
&\quad \left. - \log q\left(b_t|o_{\leq t}, a_{<t}\right) + \log p\left(o_t|b_t\right) + \log \mathcal{R}\left(r_t|b_t\right)\right]
\end{aligned} \tag{13}
$$

, where $\mathbf{b} = b_{1:T}, \mathbf{a} = a_{1:T}, \mathbf{o} = o_{1:T}$.

B.2 More result curves

Figure 5 shows the training dynamics of DreamCUB. The left plot illustrates the policy KL divergence, which reflects the difference between the current policy and the reference model. While KL naturally increases during PPO training, we keep it within a controlled range to maintain stability. The right plot shows the reward steadily increasing and eventually converging, indicating good training stability and convergence.

As shown in Figure 6, although the Acc is slightly higher when gamma is set to 1.0, the D-2 metric drops significantly. Considering both indicators, setting gamma to 0.9 achieves the best overall performance and brings out the full potential of the algorithm.

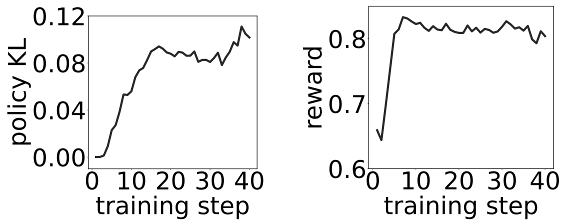

Figure 5: More training plots of DreamCUB, including the policy KL (left) and reward (right).

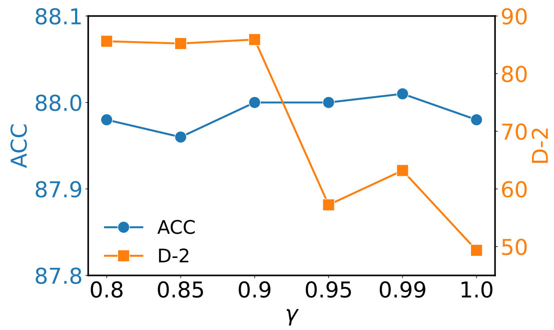

Figure 6: Curves of Acc and D-2 variations under different gamma values.

B.3 Per-emotion automatic metrics

Table 11 presents the performance of different models across four dialogue emotions. Notably, our model demonstrates a more uniform distribution of performance across different emotional categories in various metrics, thereby mitigating emotion-related bias.

B.4 Per-strategy automatic metrics

Table 12 presents the performance of different models across four dialogue emotions on the DailyDialogue dataset, using several automatic evaluation metrics. Overall, DreamCUB consistently outperforms the baselines across all metrics, demonstrating stronger generation quality and better strategic alignment.

	Model	Emotion							
		no emo	**happiness**	**surprise**	**fear**	**disgust**	**sadness**	**anger**	**total**
ACC	+ SFT	91.65	0.00	23.00	0.00	2.63	0.00	0.00	76.76
	+ COT+SFT	99.10	8.09	1.00	0.00	0.00	0.00	1.14	83.48
	+ FSM+SFT	99.81	0.62	0.00	0.00	0.00	5.26	0.00	83.28
	DreamCUB	95.65	56.61	55.00	21.43	15.79	31.58	32.95	88.05
MaF1	+ SFT	87.17	0.00	8.13	0.00	5.13	0.00	0.00	14.35
	+ COT+SFT	90.96	14.34	1.72	0.00	0.00	0.00	2.15	15.60
	+ FSM+SFT	90.89	1.23	0.00	0.00	0.00	8.99	0.00	14.44
	DreamCUB	93.17	62.81	56.70	30.00	27.27	44.44	41.73	50.88
$bias$	+ SFT	2.21	1.23	2.45	2.45	1.07	2.45	1.57	2.03
	+ COT+SFT	0.66	1.98	1.61	2.45	1.50	1.74	2.45	1.98
	+ FSM+SFT	0.78	1.99	2.45	2.45	2.45	2.45	1.79	2.22
	DreamCUB	0.65	1.52	1.05	2.45	1.42	2.45	1.07	0.74

Table 11: Per-emotion automatic metrics on DailyDialogue.

	Model	Strategy				
		directive	**inform**	**question**	**commissive**	**total**
ACC	+ SFT	1.30	78.85	47.00	74.77	60.19
	+ COT+SFT	0.37	78.02	51.88	69.91	60.11
	+ FSM+SFT	3.15	85.85	50.75	67.28	64.05
	DreamCUB	42.79	80.83	58.41	68.34	67.80
MaF1	+ SFT	2.55	75.86	44.24	56.62	44.82
	+ COT+SFT	0.74	76.01	44.67	58.19	44.90
	+ FSM+SFT	6.01	78.48	49.78	59.17	48.36
	DreamCUB	48.53	77.78	61.38	61.46	62.29
$bias$	+ SFT	0.60	0.76	0.77	0.73	0.82
	+ COT+SFT	0.60	0.76	0.77	0.73	0.82
	+ FSM+SFT	0.61	0.83	0.77	0.77	0.66
	DreamCUB	0.62	0.59	0.65	0.60	0.33
B-2	+ SFT	4.45	7.25	6.74	7.96	6.81
	+ COT+SFT	4.61	6.80	7.25	7.07	6.61
	+ FSM+SFT	6.50	5.50	7.05	4.44	5.85
	DreamCUB	10.20	12.38	12.11	9.42	11.65
R-L	+ SFT	14.59	19.92	17.00	19.72	18.54
	+ COT+SFT	14.69	19.13	17.74	18.22	18.09
	+ FSM+SFT	21.28	21.50	23.02	21.20	21.80
	DreamCUB	25.15	30.62	28.14	30.38	29.09
D-2	+ SFT	59.82	53.18	55.81	58.77	43.36
	+ COT+SFT	58.03	53.18	54.25	56.37	42.87
	+ FSM+SFT	62.07	55.83	54.10	60.59	47.43
	DreamCUB	66.25	59.24	59.15	67.77	49.36

Table 12: Per-strategy automatic metrics on DailyDialogue.