Convert Language Model into a Value-based Strategic Planner

Xiaoyu Wang[1,2*], Yue Zhao[1], Qingqing Gu[1], Zhonglin Jiang [1], Yong Chen[1], Luo Ji[1†]

[1] Geely AI Lab, Beijing, China
[2] Beijing Institute of Technology, Beijing, China
Correspondence: Luo.Ji1@geely.com

Abstract

Emotional support conversation (ESC) aims to alleviate the emotional distress of individuals through effective conversations. Although large language models (LLMs) have obtained remarkable progress on ESC, most of these studies might not define the diagram from the state model perspective, therefore providing a suboptimal solution for long-term satisfaction. To address such an issue, we leverage the Q-learning on LLMs, and propose a framework called straQ*. Our framework allows a plug-and-play LLM to bootstrap the planning during ESC, determine the optimal strategy based on long-term returns, and finally guide the LLM to response. Substantial experiments on ESC datasets suggest that straQ* outperforms many baselines, including direct inference, self-refine, chain of thought, finetuning, and finite state machines.

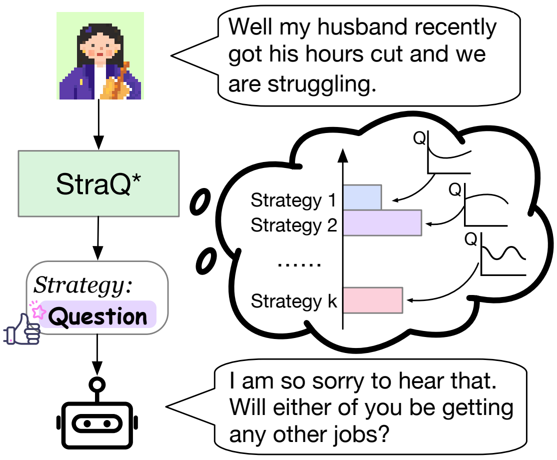

Figure 1: Paradigm of straQ*. A plug-and-play LLM-based planner selects the optimal strategy from maximized Q, then steers the LLM to enhance the response.

1 Introduction

Emotional Support Conversation (ESC) refers to dialogues aimed at alleviating a seeker's emotional distress and challenges. Effective ESC is based on relational, psychological, and physical theories (Rains et al., 2020) and has been widely explored in artificial intelligence research (Liu et al., 2021; Zhao et al., 2023). With advancements in LLMs, these models have shown strong performance in ESC (Zheng et al., 2023; Kang et al., 2024). However, most LLM-based studies focus on immediate solutions without long-term support strategies. For example, while Liu et al. (2021) defines ESC in three stages (Exploration → Comforting → Action), LLMs often struggle with smooth transitions, leading to strategy biases.

Motivated by the recent progress of reinforcement learning (RL) on LLM-based studies (Li et al., 2024b; Zhou et al., 2024; Wang et al., 2024a), we propose that ESC tasks can be defined as a strategy-level MDP, therefore value-based RL can help mitigate the aforementioned challenges. Given the current seeker's utterance, emotion and conversational history, the LLM can be prompted to identify the long-term return of strategy, learn and produce the action value, and plan the optimal strategy. The determined strategy can then be prompted to another LLM to produce improved response, guided by the strategy.

In this paper, we propose a new framework called strategic Q* (straQ*), which converts LLM into a value-based strategic planner. We use the deep Q-learning (DQN) on LLM to provide a strategic Q function, with the strategy as the textual action. We use the averaged logits of actions to denote the Q value, and update the LLM parameter by the famous Bellman equation. By this manner, we convert the next-token prediction to next-strategy prediction, bootstrapping the TD loss of strategies instead of the original cross-entropy loss. This Q-net is used as a plug-and-play strategic planner, along with the conversation LLM to produce the

[*] Work was done during the internship at Geely.
[†] Corresponding Author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1444–1456

ultimate response. Our main contributions can be summarized as follows:
(1) We define the strategy-level MDP, and formulate the LLM architecture as a Q-function with textual input of state and strategy.
(2) We empirically verify that pretrained LLM can be finetuned by Bellman Equation and converges to optimum returns, with the averaged logit of action tokens as the q-value.
(3) Substantial experiments on ESConv and EmpatheticDialogues indicate that straQ* results in higher response quality and more reasonable planning of strategies.
(4) We design two reward mechanisms including imitation and distillation, with the former better at automatic metrics, while the latter better at human scoring and generalization.

2 Preliminary

Strategy-level MDP. The Markov decision process (MDP) is usually defined as a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, $\mathcal{R}$ is the reward set, γ is the discounting factor of rewards, and $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the state transition function. In this work, we formalize the ESC task as a strategy-level MDP, with the action space $\mathcal{A} = \{a\}$ as the set of possible strategies.

Q-Learning. In value-based RL, the goal is to learn the state-value function $V(s)$ or the state-action value function $Q(s, a)$, such that the determined action achieves the highest expected discounted cumulative reward:

$$a^{\star} = \arg\max_{a} Q(s, a) \leftarrow \arg\max \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

which is solved by the famous Bellman Equation:

$$Q^*(s, a) = r(s, a) + \gamma \max_{a'} Q^*(s', a') \quad (1)$$

in which the superscript $'$ indicates the next step. Instead of explicitly implementing the above equation, Deep Q-learning (DQN) approximates the maximization of the right-hand side with the deep value networks:

$$\mathcal{L}(\theta) = |r(s, a) + Q_{\phi}(s', a') - Q_{\theta}(s, a)|^2 \quad (2)$$

where $\mathcal{L}$ is the loss, θ and ϕ are parameters of the Q-net and the target Q-net, respectively. ϕ can be periodically synchronized from θ.

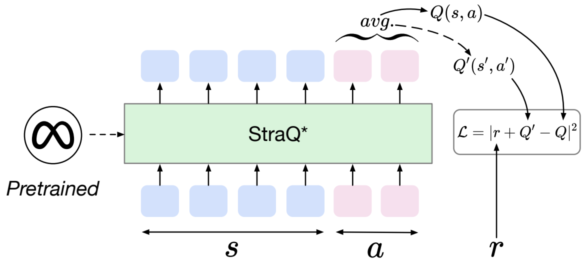

Figure 2: The training framework of straQ*. Averaged log probability of action tokens is defined as $Q(s, a)$ which deduces the training loss from Bellman Equation.

3 Methodology

3.1 Task Definition

The problem of emotional-support conversation (ESC) can be characterized by an interleaved sequence of seeker $query$ and supporter $response$. To strengthen emotional-support performance, recent studies (Liu et al., 2021; Rashkin et al., 2019) enhance the data content by augmenting the set of support strategies $\mathcal{A}$ and seeker emotions $\mathcal{E}$. For each conversation session, the background $description$ is also annotated on the session-level. Such augmented ESC can then be described as

$$\begin{aligned} & desc, \{query(t), e(t), a(t), resp(t)\}_{0:T} \\ & a \in \mathcal{A}, \quad e \in \mathcal{E} \end{aligned} \quad (3)$$

in which $desp$ and $resp$ are the abbreviations of $descrition$ and $response$, and T is the total number of conversation turns. At turn t, we denote the conversation history as

$$h(t) = \{query(t), e(t), a(t), resp(t)\}_{0:T-1} \quad (4)$$

Then the ESC sample at time t can be alternatively expressed as $\{h(t), query(t), e(t), a(t), resp(t)\}$.

3.2 System Variables

We define important system variables as follows:
• **State:** The state is a combination of description, emotion, history and query, *i.e.*, $s = \{desc, e, h, query\} \in \mathcal{S}$.
• **Action:** The conversational strategy, $a \in \mathcal{A}$.
• **Reward:** The reward r_t can be viewed as the instantaneous satisfaction of the seeker, which can be either inferred from an annotated datasets, or generated by an off-the-shelf model evaluator (LLM-as-the-Judge).
• $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function. After

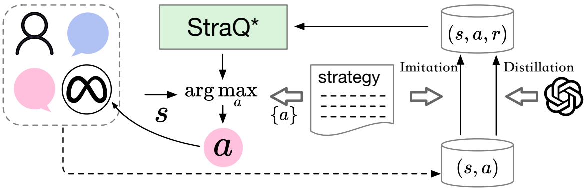

Figure 3: Diagram of our pipeline. Reward is annotated by the imitation or distillation strategy. The action is chosen from the strategy which maximize the Q value.

the t-th turn, h is updated by appending the current $query(t)$ and $resp(t)$, the seeker reacts further with new $query(t+1)$ and $e(t+1)$, and the step is incremented by one.

3.3 Implementation on Language Models

LLM-based value function. Our implementation starts from a pretrained LLM, with the parameter of θ. We assume there is an instruction template with the placeholder of s, denoted by $\mathcal{I}(s)$. This instruction can be concatenated with a, $\mathcal{I} \oplus a$. Both state and state-action values can be obtained from the semantic understanding of LLM:

$$Q_\theta(s,a) \leftarrowtail \text{LLM}_\theta(\mathcal{I}(s) \oplus a) \tag{5}$$

where $\leftarrowtail$ means to average the action logits.

Training a strategic value-function. By replacing the Q-net in Equation 2 by the above expressions, we finetune the LLM by the Bellman Equation loss on last token logit. As in standard language modeling, the causal masking of the transformer allows us to perform Bellman updates on entire sequences in parallel. Figure 2 exhibits this training framework.

We keep the setting of the target Q-net, which is the same LLM architecture, while its parameter ϕ are periodically synchronized from θ.

Inference the optimal strategy. Instead of decoding the next token, the finetuned LLM produces logits of available strategies, and the optimal strategy can be determined from the maximum logit

$$a^\star \leftarrow \arg\max \text{LLM}(\mathcal{I}(s) \oplus a), a \in \mathcal{A} \tag{6}$$

Instruction template. We briefly exhibit our instruction $\mathcal{I}(s)$ here:

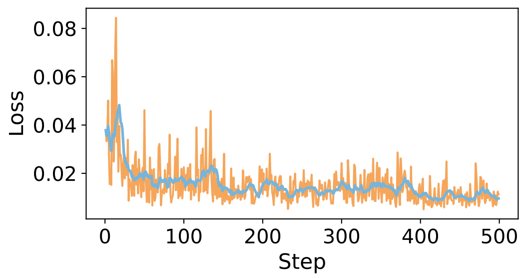

Figure 4: Training loss curve of straQ*.

Interaction Summary

Description: $\{desc\}$ **User's emotion:** $\{e\}$
History: $\{h\}$ **Query:** $\{query\}$
Please select the best strategy:
(1) $\{strategy1\}$ **(2)** $\cdots$ **(K)** $\{strategyK\}$

with the full version in Appendix A.2. By forming the prompt To further strengthen the understanding capability of LLM on the strategy selection, we formulate $\mathcal{I}$ as a multi-choice question (MCQ), instead of a plain question, forcing the LLM to choose one of the option numbers. Accordingly, the action set becomes the set of possible strategy index $a \in \mathcal{A} := \{1, 2, \cdots, K\}$ where K is the total number of strategies.

3.4 Reward Definitions

Choice of rewards may be crucial especially when the sampling is constrained by an offline dataset. In this paper, we study two reward mechanisms:
(1) Distillation: for each (s, a) pair from the dataset, we let a strong-basis LLM (*e.g.*, GPT-4) to provide a judge score from 0 to 5 (The detailed judge prompt is in Appendix A.2). Since this manner distills the knowledge from a teacher model, we call this variant as **straQ*-distill**.
(2) Imitation: we consider each (s, a) pair from the dataset is an expert demonstration, therefore, always assigned with r of $+1$. To amplify the distribution, we randomly sample a different a and assign with r of -1. The positive-negative ratio is 1:1. Since this manner imitates the positive samples directly, this variant is called **straQ*-imit**.

Figure 3 shows the entire pipeline of straQ*.

4 Experiment

4.1 Setting

Implementation. Llama3.2-1B-instruct (AI@Meta, 2024) is employed as the base model. Training is conducted on OpenRLHF (Hu et al.,

Strategies	Abbr.	Stage
Question	**Que.**	I
Restatement or Paraphrasing	**Res.& Par.**	I
Reflection of Feelings	**Ref.**	II
Self-disclosure	**Self-Dis.**	II
Affirmation and Reassurance	**Aff.& Rea.**	III
Providing Suggestions	**Pro.**	III
Information	**Inf.**	III
Others	**Others**	-

Table 1: Strategy names, abbreviations and stages.

2024), with the learning rate of $5.0e-6$, window length of 2048, batch size of 64, and epoch of 4. The target network update frequency is set to 10, the replay buffer size is 12,000, and $\gamma = 0.85$.

Datasets. Training of straQ* requires the annotation of strategies. We use ESConv (Liu et al., 2021) as the training set and also the in-domain (ID) test. ESConv provides $K = 8$ strategies, which belong to three ESC stages: *Exploration (I), Comforting (II)* and *Action (III)*. Table 1 shows their full names, abbreviations and corresponding stages.

Furthermore, EmpatheticDialogues (Rashkin et al., 2019) is employed as the out-of-domain (OOD) evaluation, since EmpatheticDialogues does not have the strategy annotation. For the ID test, both strategy-related and response-related results can be provided. For the OOD test, only zero-shot response-related results are provided. Appendix A.1 provides a more detailed introduction of ESConv and EmpatheticDialogues.

4.2 Evaluation Methods

Automatic Metrics. To evaluate the quality of strategy determination, we refer the evaluation methods proposed by Kang et al. (2024), which uses **proficiency** $\mathcal{Q}$ based on macro-F1, and **preference bias** $\mathcal{B}$ based on Bradley-Terry model (Bradley and Terry, 1952). Smaller $\mathcal{B}$ means less bias, therefore is better. We also include the strategy prediction accuracy (Acc). For response quality, we utilize the famous Bleu-2 (B-2), Rouge-L (R-L), Distinct-2 (D-2) and CIDEr, calculating from the similarity with the ground truth response.

Human Scoring. Similar with Kang et al. (2024), we annotate with the dimensions of *Acceptance*, *Effectiveness*, *Sensitivity*, *Fluency*, and *Emotion*, and the ultimate purpose, seeker's *Satisfaction*.

Baselines. We consider the following baselines:
(1) Direct: directly inference the LLM.
(2) Direct-Refine: the model immediately revises its response within the same utterance to incorporate emotional support considerations.
(3) Self-Refine (Madaan et al., 2023): the model considers the emotional support, generates a feedback from the initial response, then refines the response based on the feedback.
(4) CoT (Wei et al., 2022): steered by the chain-of-thought prompt, the model first identifies *emotion*, then generates *strategy*, and finally *response*.
(5) FSM (Wang et al., 2024b): the finite state machine with finite sets of states and state-transitions triggered by inputs, and associated discrete actions.

Methods	Acc ↑	$\mathcal{Q}$ ↑	$\mathcal{B}$ ↓	B-2 ↑	R-L ↑
LLaMA3-8B-Instruct					
Direct	11.80	10.26	1.61	3.47	10.64
+ Direct-Refine	17.08	11.07	1.27	3.10	6.13
+ Self-Refine	17.58	13.61	1.92	3.34	9.71
+ CoT	15.32	10.38	1.69	3.16	10.50
+ FSM	17.37	11.15	0.81	**4.12**	<u>11.83</u>
+ 1B straQ*-distill (ours)	<u>41.22</u>	<u>38.95</u>	**0.57**	<u>3.89</u>	11.80
+ 1B straQ*-imit (ours)	**46.83**	**43.15**	<u>0.80</u>	<u>3.89</u>	**12.84**
LLaMA3-8B-Instruct + SFT					
Direct	32.43	21.29	1.28	6.97	16.59
+ CoT	30.80	17.70	1.35	6.51	15.00
+ FSM	28.83	18.36	1.32	<u>7.57</u>	**17.42**
+ 1B straQ*-distill (ours)	<u>41.22</u>	<u>38.95</u>	**0.57**	7.01	16.93
+ 1B straQ*-imit (ours)	**46.83**	**43.15**	<u>0.80</u>	**7.63**	<u>17.30</u>

Table 2: ID Results of automatic metrics including Acc, $\mathcal{Q}$, $\mathcal{B}$, Bleu-2 (B-2) and Rouge-L (R-L) on the testset of ESConv. The best results of each LLMs are **bolded** and the second best are <u>underlined</u>.

Methods	B-2	R-L	Dist-2	CIDEr
Direct	3.09	9.91	25.23	1.60
+ CoT	2.91	9.79	32.65	1.37
+ FSM	3.33	10.80	33.37	2.96
+ 1B straQ*-distill (ours)	**4.49**	**12.93**	<u>46.53</u>	**8.36**
+ 1B straQ*-imit (ours)	<u>4.27</u>	<u>12.66</u>	**46.80**	<u>8.11</u>

Table 3: OOD finetuned results of Bleu-2 (B-2) and Rouge-L (R-L) on EmpatheticDialogues. The best results of each LLMs are **bolded** and the second best are <u>underlined</u>.

4.3 Results

Training Curves. Figure 4 shows the training loss curve of straQ*for 500 steps (approximately 3 epochs). Although the loss initially fluctuates significantly, it adapts to the new training paradigm, and finally tends to be stable.

Automatic Evaluations. Table 2 presents the automatic metrics on the ID evaluation, with the basis of either the original LLM, or the specifi-

Method	Human Annotation						
	Fluency	Emotion	Acceptance	Effectiveness	Sensitivity	Alignment	Satisfaction
Original dataset	3.51	3.61	3.40	3.10	3.50	3.20	3.30
Llama3-8B-Instruct	2.95	3.00	2.60	2.40	2.70	2.70	2.60
+ Direct-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
Llama3-8B-Instruct+ SFT	3.15	3.40	2.70	2.70	2.90	3.30	2.90
+ CoT	**3.67**	3.61	3.22	3.67	3.56	3.35	3.45
+ straQ*-distill (ours)	3.52	**3.65**	**3.59**	**3.73**	**3.71**	**3.62**	**3.66**
+ straQ*-imit (ours)	3.42	3.25	3.23	3.07	3.10	3.21	3.13

Table 4: Averaged Human evaluation of response quality on ESConv and EmpatheticDialogues.

cally finetuned version. Compared to baselines, straQ*generally achieves higher strategy accuracy, lower bias, and higher similarity to the ground truth responses. Furthermore, straQ*-imit performs better than straQ*-distill on this setting, suggesting that the imitation-version of rewards result in better ID performance.

In Table 3, we further compare the OOD results of the models finetuned by ESConv, with the strategy lists inferred from ESConv. Results suggest that straQ*demonstrates strong generalization than these baselines. Specifically, straQ*-distill surpasses straQ*-imit this time, indicating the distilled knowledge from the teacher model is more general than simply imitating a limited dataset.

Human Evaluation. The results of the crowd-sourcing evaluation shown in Table 4 indicate that straQ*-distill outperforms the baseline methods in various metrics, such as Fluency, Emotion, and Satisfaction. It also performs better than the replies in the source data. Conversely, straQ*-imit is slightly lower than the source data in performance. Using the GPT-4 score as reward, straQ* can determine strategies more from the aspect of performance optimization, not simply imitating the demonstration.

Ablation Study. Two ablations are studied:
(1) *w/ value head*: append the model with a classification head which produces the score logit.
(2) *auto-regressive*: keep the cross-entropy loss with the ground truth action as the target text.
In more detail, *w/ value head* is the usual solution for a reward model in RLHF, while *auto-regressive* can be viewed as a standard fine-tuning solution for the strategic planner. Table 6 shows that straQ*outperforms both of them in various automatic metrics, indicating our methodology can better align with the strategy semantics and more accurately capture the strategic value.

Sensitivity Analysis. Figure 6 shows the ID performance evolutions on different γ choices. Smaller γ means we are more focused on the transient performance and relatively neglect the long-term value. Results show that the optimal accuracy of strategy happens on $\gamma = 0.9$, while the best response-related metrics correspond to $\gamma = 0.85$. Because B-2 and R-L are similarity-based, the current reward is more relative to them than future rewards. Therefore, this observation is reasonable.

4.4 Discussions

Scalability and application. Figure 6 (bottom-right) also compares the B-2 results on different model sizes. As the model becomes larger, the performance also increases, indicating straQ* can have good scalability. However, larger models result in higher computation overhead and slower speed, which may hinder the practical application of straQ*. Therefore, in the formal application, we still adhere to the 1B choice, employing it as a lightweight planner.

From previous results, we utilize straQ*-distill in the actual application to have better generalization and better alignment with human knowledge.

Returns of Strategies. Table 5 further analyzes two important indicators of value-based RL, the averaged rewards and values. In this analysis, the rewards are provided by GPT-4. straQ* achieve both higher <reward> and <value> than direct inference of the base model, as well as the annotation of original dataset. This result shows that straQ* statistically obtains higher returns, which is the primary purpose of Q-Learning.

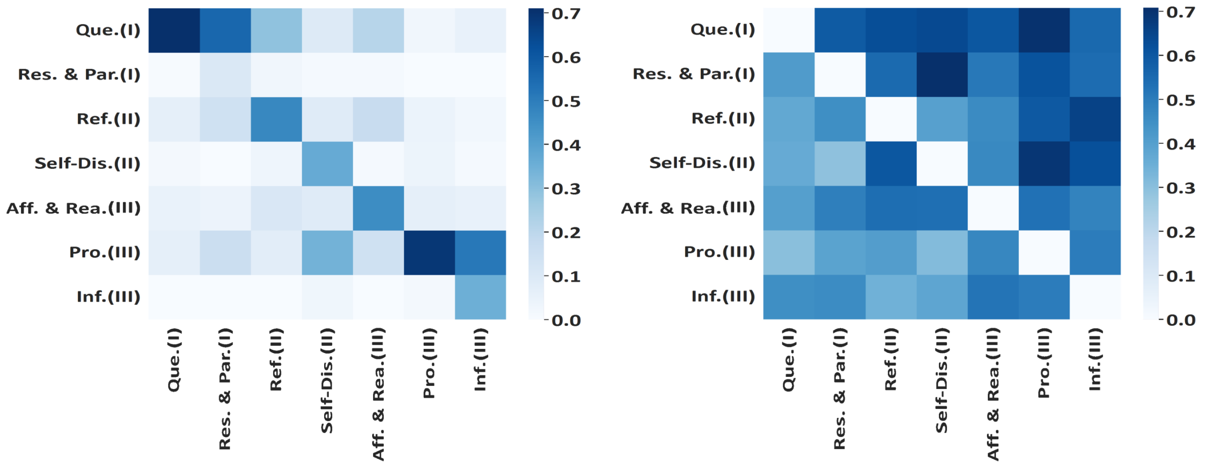

Figure 5: Distribution of strategy determined by straQ*. Strategies are labeled with the stage index (I, II, III) which represents the general scenario Exploration → Comforting → Action in ESC. Left: the confusion matrix (acted strategy (row) VS ground truth strategy (column)). Right: the transition matrix (acted strategy (row) VS the next-acted strategy (column)).

Method	<reward>	<value>
Original dataset	3.01	252.09
Llama3-8B-Instruct	3.66	346.31
straQ*-distill (ours)	**3.99**	424.78
straQ*-imit (ours)	3.72	**445.95**

Table 5: Average reward (eval by GPT-4) of strategy determination on the testset of ESConv.

Method	Acc ↑	$\mathcal{Q}$ ↑	$\mathcal{B}$ ↓	B-2 ↑	R-L ↑
w/ value head	19.81	11.40	1.66	6.74	15.99
auto-regressive	46.22	43.01	**0.69**	7.25	16.48
straQ*-imit	**46.83**	**43.15**	0.80	**7.63**	**17.03**

Table 6: Ablation study of straQ*-imit on ESConv.

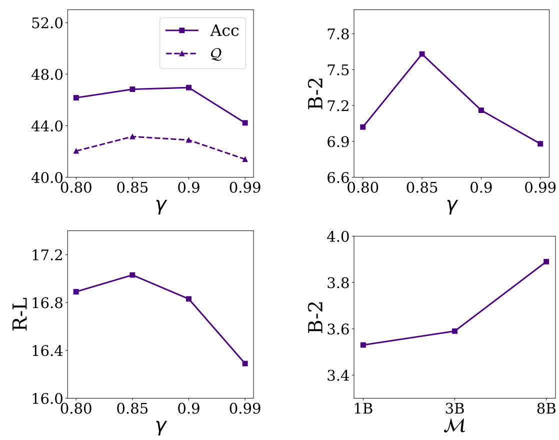

Figure 6: Sensitivity plots of straQ*-imit on different γ and model sizes. Metrics include Acc, $\mathcal{Q}$, B2, and R-L.

Strategy Prediction and Transitions. Figure 5 (Left) exhibits the confusion matrix of strategies, with the rows representing the prediction, and the columns representing the ground truth. Results show that most occurrences happen on the diagonal grids, verifying the prediction accuracy.

Figure 5 (Right) visualizes the transition matrix. A grid (i, j) means the strategy i to the strategy j, where the strategies are sorted from their ESC stages (from I to III) for both rows and columns. Therefore, transitions from an early stage to a later stage should occur on the upper-triangle region of the transition matrix. Results in the figure validate this proposition.

Detailed results of Strategies. Strategies' popularities and occurrences may differ in nature. For example, straightforward strategies like *"Question"* and *"Providing Suggestions"* may be more convenient to learn and apply. A reasonable strategic planner should reflect this frequency difference, but also avoid significant bias (*e.g.*, always determines the most frequent strategy). To further study the strategy-related performance, we further exhibit the per-strategy results in Table 8, with the comparison between straQ*-distill and SFT. One can observe that straQ* model is more balanced across different strategies, and SFT even have zero predictions on the strategy of "Information".

Typical Case. Table 7 presents a typical case of straQ*-distill in the third turn, comparing to Direct, and also the original response in the dataset. To better illustrate the effectiveness of the value-learning, we present the reward and value scores

Characters	Strategy	Response	Reward	Value
supporter	-	*Hi! Hope you are doing well?*	-	-
seeker	-	*Hi, yes am doing well*	-	-
supporter	-	*How can I help you today?*	-	-
seeker	-	*I had a drinking challenge which has been affecting my marriage for quit sometime now*	-	-
supporter	*Res.& Par. (I)*	*So as far as I understand that you had issues with your wife due to your drinking. Am I right?*	3	7.6
Direct	Res.& Par. (I)	So, it sounds like there's been a bit of tension in your marriage lately due to your drinking habits, and it even led to a confrontation that's caused a rift between you and your wife. Can you tell me more about what's been going on with your drinking?	4	8.0
straQ*-distill	**Res.& Par. (I)**	**It sounds like your drinking habits have been causing difficulties in your marriage for a while.**	3	**10.1**
seeker	-	*Yes that is true*	-	-
supporter	*Aff.& Rea. (II)*	*That is really a serious problem. I know it must be very tough time for you. Hope you are trying your best to save your marriage.*	2	5.4
Direct	Res.& Par. (I)	Just to make sure I understand correctly, it sounds like there was a confrontation with your wife about your drinking habits, and now you two aren't speaking. Is that a fair summary of what happened?	3	4.7
straQ*-distill	**Ref. (II)**	**That must be really difficult for you. I can imagine how much stress and worry this situation has brought into your life.**	4	**8.3**
seeker	-	*Yes I have really reduced my drinking to ensure I save my marriage*	-	-
supporter	*Ref. (II)*	*That's is a great initiative. I know it is tough for you but not impossible. I can feel how hard you are trying t save your relationship. I appreciate you for that.*	4	4.0
Direct	Res.& Par. (I)	Just to make sure I understand correctly, you've been trying to cut down on your drinking to address your wife's concerns, but despite your efforts, the issue still seems to be causing tension in your marriage, and now you're not talking to each other. Is that a correct summary?	2	2.0
straQ*-distill	**Aff.& Rea. (III)**	**That's a great step forward! It takes a lot of strength and commitment to make such a change, and you should be proud of yourself.**	5	**5.0**

Table 7: Case of straQ* in ESConv.

Table 8: Per-strategy automatic metrics on ESConv.

	Strategy	Acc ↑	$\mathcal{Q}$ ↑	$\mathcal{B}$ ↓	B-2 ↑	R-L ↑	Dist-2	CIDEr
SFT	Que.	57.52	**48.24**	1.60	**9.37**	**21.88**	64.42	**34.22**
	Res.& Par.	18.52	20.55	0.84	7.96	16.99	77.38	21.12
	Ref.	1.57	2.92	0.08	5.74	14.90	73.44	11.72
	Self-Dis.	2.36	4.38	**0.06**	4.99	12.30	76.71	7.85
	Aff.& Rea.	20.09	22.45	0.83	5.94	15.17	70.72	13.20
	Pro.	**75.22**	40.95	4.12	5.99	14.26	71.26	11.04
	Inf.	0.00	0.00	0.00	5.93	12.24	**78.54**	12.95
	Others	23.20	30.77	0.46	8.38	18.21	74.08	27.78
straQ*-distill	Que.	**71.07**	**60.59**	2.43	**9.48**	**22.05**	64.44	**33.50**
	Res.& Par.	8.97	16.67	**0.16**	8.45	17.30	79.05	21.60
	Ref.	40.88	38.69	0.43	5.18	13.62	75.95	9.78
	Self-Dis.	29.85	41.75	0.46	4.90	12.70	76.00	6.01
	Aff.& Rea.	36.63	42.11	0.47	6.30	15.46	70.35	13.66
	Pro.	69.08	56.13	0.60	5.95	14.03	70.51	10.62
	Inf.	26.67	40.14	0.47	5.33	13.21	**79.94**	8.11
	Others	45.93	45.95	0.53	6.60	15.19	71.85	24.34

for each response. In this case, straQ* does not simply maximize the immediate reward, but maximizes the long-term return (*i.e.*, the value), which is calculated from subsequent turns. Also, straQ* in this case exhibits a perfect stage-turnover, guiding the conversation from the first stage (strategy Res.&Par.), then the second stage (strategy Ref.), to the third stage (strategy Aff.& Rea.). Comparing to Direct (stays in I) the original response (I to II to II), planning of straQ* is more consistent with the theory proposed in (Liu et al., 2021).

5 Related Work

There are some RL studies in goal-oriented conversations (Li et al., 2024b; Zhou et al., 2024; Li et al., 2024a). For example, DAT (Li et al., 2024b) defines dialogue action tags, and then generates responses by multi-turn planning. ArCHer (Zhou et al., 2024) proposes a hierarchical RL algorithm to improve the efficiency and performance of LLMs. These works adapt the conversational LLM, and rely on ground truth annotations. In contrast, our straQ* implements an explicit, lightweight and plug-and-play planner, which balances the foundation capability and the strategic thinking.

6 Conclusion

In this paper, based on Q-learning, we propose a method named straQ*that optimizes long-term returns in emotional support conversation scenarios. Our implementation behaves as a plug-and-play strategic planner which steers the subsequent response generation. We propose two reward mechanisms, straQ*-imit and straQ*-distill, in which the former has higher automatic evaluation results, and the latter performs better in generalization and human preference alignment.

7 Limitation

There are still some limitations of straQ*. The results of human evaluation may be biased, or deviate from the judgments of actual help - seekers due to the awareness of being engaged in scoring. Then, the testset may be small. Although it has little impact on the comparison between automated and human evaluations, sample sizes for some subcategories may be insufficient when conducting a detailed analysis.

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.

Clara E Hill. 2009. *Helping Skills: Facilitating, Exploration, Insight, and Action*. American Psychological Association.

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*.

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.

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

Kenneth Li, Yiming Wang, Fernanda Viégas, and Martin Wattenberg. 2024b. Dialogue Action Tokens: Steering Language Models in Goal-Directed Dialogue with a Multi-Turn Planner. *Preprint*, arXiv:2406.11978.

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.

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.

Stephen A Rains, Corey A Pavlich, Bethany Lutovsky, Eric Tsetsi, and Anjali Ashtaputre. 2020. Support seeker expectations, support message quality, and supportive interaction processes and outcomes: The case of the comforting computer program revisited. *Journal of Social and Personal Relationships*, 37(2):647–666.

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.

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.

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

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.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. TransESC: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739, Toronto, Canada. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

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

A Further Implementation Details

A.1 Dataset details

Emotional Support Conversations. Emotional Support Conversation (ESC) is a task aimed at alleviating users' negative emotions (e.g., anxiety, depression), where **supporters** assist **seekers** in managing emotions triggered by issues like work crises or interpersonal conflicts. Unlike emotion recognition tasks, ESC integrates psychological counseling mechanisms into the dialogue generation process, offering a deeper, context-sensitive solution for emotion regulation. The ESC dataset generally has the following attributes:

- **Emotion**: Including emotion types and intensities, which help accurately capture the psychological state of the help-seeker.
- **Help-seeker profile**: A brief survey before each conversation that provides insight into the current situation of the help-seeker, revealing the challenges they are facing.
- **Situation**: A brief survey before each conversation that provides insight into the current situation of the help-seeker, revealing the challenges they are facing.
- **Strategy**: The response rule selected for the current turn based on the seeker's emotional state. There are eight predefined rules in total.
- **Response**: The supporter's response generated based on the history, inferred state, and selected rule.

In this paper, we mainly study two typical ESC datasets, ESConv and EmpatheticDialogues. ESConv has exactly the aforementioned architecture while EmpatheticDialogues lacks the Strategy. Table 9 summarizes the basic statistical information of ESConv and EmpatheticDialogues.

Category		ESConv	EmpatheticDialogues
# Sessions		1.3K	2.5K
# Uttr		38K	11.0K
Average # Uttr		28.9	4.3
Average Uttr Len		18.8	16.7
Seeker	# Uttr	20K	5.7K
	Avg # Uttr	15.4	2.2
	Avg Uttr Len	16.8	20.8
	# Emotions	11	32
Supporter	# Uttr	18K	5.2K
	Avg # Uttr	13.6	2.1
	Avg Uttr Len	21.0	12.3
	# Strategies	8	-

Table 9: Statistics of ESConv and EmpatheticDialogues. 'Uttr' abbreviates Utterance.

Emotion Type	# Occurrence
anger	111
anxiety	354
depression	334
disgust	40
fear	95
nervousness	13
sadness	308
shame	42

Table 10: Emotion statistics of ESConv.

ESConv. Motivated by the Helping Skills Theory (Hill, 2009), Liu et al. (2021) divides ESC into three sequential stages: *Exploration*, *Comforting*, and *Action*, and proposes a dataset called ESConv. For each sample, the conversation is multi-turn, with the dialogue background and user emotion annotated. Upon each utterance of the supporter, 8 distinct support strategies are annotated. Table 11 exhibits an example of ESConv.

Table 12 provides definitions of support strategies in ESConv. Table 10 lists the emotion types and their occurrences in the dataset. The emotion types include anger, anxiety, depression, disgust, fear, nervousness, sadness, and shame.

EmpatheticDialogues. EmpatheticDialogues (Rashkin et al., 2019) is a dataset that consists of empathetic conversations. It aims to help in the development of empathetic language models by providing a large number of dialogues that express empathy.

A.2 Prompt format

Instruction template. To further strengthen the understanding capability of LLM on the strategy selection, we define the instruction as a multi-choice question (MCQ), forcing the LLM to choose one of the option numbers, instead of a plain question. Below is the content of the instruction template $\mathcal{I}$:

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 11: An example of *ESconv*.

Strategies	Abbr.	Definitions
Question	**Que.**	Inquiring about problem-related information to help the seeker clarify their issues, using open-ended questions for best results and closed questions for specific details.
Restatement or Paraphrasing	**Res.& Par.**	A simple, more concise rephrasing of the help-seeker's statements that could help them see their situation more clearly.
Reflection of Feelings	**Ref.**	Articulate and describe the help-seeker's feelings.
Self-disclosure	**Self-Dis.**	Divulge similar experiences that you have had or emotions that you share with the help-seeker to express your empathy.
Affirmation and Reassurance	**Aff.& Rea.**	Affirm the help seeker's strengths, motivation, and capabilities and provide reassurance and encouragement.
Providing Suggestions	**Pro.**	Provide suggestions about how to change, but be careful to not overstep and tell them what to do.
Information	**Inf.**	Provide useful information to the help-seeker, for example with data, facts, opinions, resources, or by answering questions.
Others	**Others**	Exchange pleasantries and use other support strategies that do not fall into the above categories.

Table 12: Strategy names, abbreviations and detailed definitions in ESConv.

You are a psychological consultant providing support to a seeker. The seeker's basic situation is as follows:
Emotion: $\{e\}$
Description: $\{desp\}$
Below is the conversation history between the seeker and the supporter:
$\{h\}$
The seeker's current query is:
$\{query\}$
Based on the above context, please select the most appropriate response strategy from the following options:
strategy #(1) $\{a_1\}$
...
strategy #(k) $\{a_k\}$
Please provide your selection in the format of (1) through (k). Your selection is:

Generation prompt. Below is the prompt used by the conversational foundation LLM for the response generation:

You are a psychological consultant providing support to a seeker. The seeker's basic situation is as follows:
Emotion: $\{e\}$
Description: $\{desp\}$
Below is the conversation history between the seeker and the supporter:
$\{h\}$
The seeker's current query is:
$\{query\}$
The current response strategy is:
$\{a\}$
Based on the current response strategy and other information, please act as a supporter and provide the best response. Keep replies brief without additional pronouns or extra elements.

Prompt of GPT-4 for reward generation. Below is our prompt of GPT-4 to generate the rewards for straQ*-distill:

A.3 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 *sat-*

You are a psychological consultant providing support to a seeker. The seeker's basic situation is as follows: Emotion: $\{e\}$ Description: $\{desp\}$ Below is the conversation history between the seeker and the supporter: $\{h\}$ The seeker's current query is: $\{query\}$ Please evaluate whether the response is appropriate: $\{resp\}$ Based on the information above, evaluate whether the response is suitable. Please remember to respond with a single integer number from 1 to 5, where 1 indicates "not suitable" and 5 indicates "very suitable". Please also provide a brief explanation of your decision.

Table 13: Template of GPT-4 scoring.

isfaction. 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.
To guarantee the accuracy and reliability of the evaluation results, a pre - evaluation training program is meticulously designed and implemented. During this training, the evaluation criteria are clearly and systematically expounded. Moreover, detailed explanations and scoring rules corresponding to each score are provided.
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 Scoring details of GPT-4

Table 14 presents GPT-4 score statistics across different response strategies. The overall average score is 3.67, with a median of 4. The most frequently used strategies are Others (17.8%), Questioning (17.6%), and Affirmation & Reasoning (16.7%), while Restating & Paraphrasing (6.7%) and Information Providing (6.8%) appear less often. In terms of average score, Providing Opinions, Others, and Affirmation & Reasoning score the highest (all around 3.76–3.77), whereas Restating & Paraphrasing and Self-Disclosure have the lowest average scores (3.48).

Strategy	Count	Ratio	Max	Min	Avg	Median
Que.	2574	17.6%	5	1	3.54	4
Res.& Par.	981	6.7%	5	1	3.48	3
Ref.	1253	8.6%	5	2	3.65	4
Self-Dis.	1410	9.6%	5	2	3.48	3
Aff.& Rea.	2444	16.7%	5	1	3.76	4
Pro.	2367	16.2%	5	1	3.77	4
Inf.	995	6.8%	5	2	3.75	4
Others	2600	17.8%	5	1	3.77	4
Total	14624	100.0%	5	1	3.67	4

Table 14: Statistics of GPT-4 score.