Enlarged Strategy Search with In-context Rewarding for Emotional Support Conversation

Anonymous ACL submission

Abstract

Emotional support conversation is an important task, which aims to provide safe and effective support for emotional seekers. Recent approaches typically employ multi-agent collaborative reasoning to identify appropriate strategies for response generation. However, their strategy space is often small and ambiguous, which leads to suboptimal decisions in real-world cases, while a large fine-grained space would be more challenging for agents to select the appropriate one. We propose a novel framework, Enlarged Strategy Search with In-context Rewarding (**EASrIeR**) for emotional support conversation. We use a turn-aware MCTS-style tree search procedure that identifies high-quality fine-grained strategy candidates from a large deliberated space for downstream response generation and selection, guided by an LLM-as-judge composite reward. While our fine-grained strategy space better captures real-world needs, the turn-aware constraint limits the search space to a smaller range, and the MCTS-style search generates and selects promising candidate strategy-response pairs for final response generation. During search, our reward combines LLM-as-judge reward with retrieved experience cues, which provide effective guidance without additional training. Experiments on the ESConv benchmark show that EASrIeR improves automatic generation metrics over competitive baselines.

1 Introduction

Mental-health AI systems are increasingly useful for providing timely assistance (Lai et al., 2023; Xu et al., 2024). Emotional support conversation (ESC) task (Liu et al., 2021) aims to help users cope with negative life events through supportive dialogue. An ESC system should recognize the user's underlying concern, express empathy, and provide appropriate help. As large language models (LLMs) have shown strong great in-context learning ability (Brown et al., 2020), recent work widely explore their ability for ESC (Xu et al., 2025b; Ye et al., 2025; Zhang et al., 2024; Madani and Srihari, 2025; Tu et al., 2025; Xu et al., 2025a).

Since it is difficult for LLMs to directly generate high-quality responses, many ESC methods first select a *support strategy*, then generate a strategy-conditioned response. For example, one of the most widely used ESC benchmarks, ESConv (Liu et al., 2021), defines eight support strategies and annotates supporter utterances. Mixed-Initiative (Chen et al., 2023) shows that well-designed prompting can steer controllable dialogue generation. Some studies use LLM-generated support data or teacher-style supervision to improve emotional support models (Zheng et al., 2024). However, some studies show that LLMs exhibit preference bias toward a narrow subset of support strategies, which can hurt balanced emotional support generation (Kang et al., 2024; Zhao et al., 2023). MultiAgentESC (Xu et al., 2025a) addresses this issue by moving from strategy-centered generation to response-centered multi-agent comparison over multiple strategy-conditioned responses. However, it is limited to a relatively small strategy space, and candidate selection may still inherit the preference bias of underlying LLMs.

To address these issues, we present Enlarged Strategy Search with In-context Rewarding (**EASrIeR**). EASrIeR uses a dialogue-turn-aware MCTS-style strategy search procedure to identify high-quality strategy-conditioned candidates from a finer-grained strategy space for downstream response generation. Based on the 8 strategies in ESConv as parent strategies, we construct a larger fine-grained strategy space to capture diverse real-world support needs. To mitigate LLMs' bias, we use a turn-aware constraint to limit the strategy space, and use the Monte Carlo tree search (MCTS) method to generate and select strategy-constrained responses. During search, our reward uses retrieved experience as prior and LLMs' in-context judg-

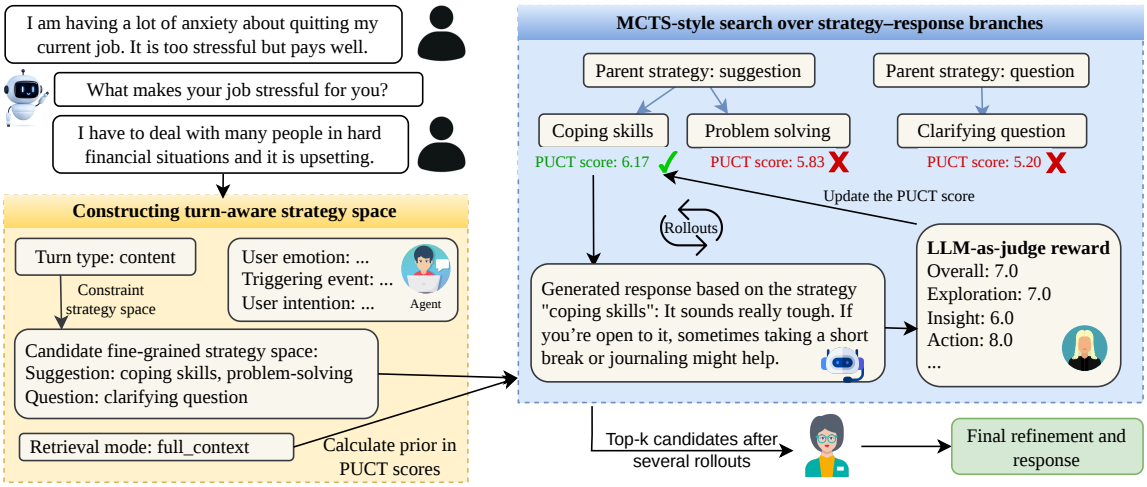

Figure 1: An overview of the proposed EASRIER.

ment scores, providing effective guidance without additional training. We evaluate EASRIER on the ESConv benchmark and show that it outperforms competitive baselines.

Our main contributions are summarized as follows: ❶ We propose a new method EASRIER for fine-grained strategy-conditioned response search in emotional support conversation. ❷ We introduce an in-context composite reward that provides effective guidance for candidate evaluation. ❸ Experiments on the benchmark demonstrate the effectiveness of EASRIER over diverse baselines.

2 The Proposed EASRIER

Problem Definition. Emotional support conversation (ESC) (Liu et al., 2021) requires generating a supportive response that helps a seeker regulate emotions and cope with stressors. Given a dialogue history $H = \{(u_i, r_i)\}_{i=1}^{t-1}$ with seeker utterances u_i and supporter utterances r_i, the system outputs the next supporter response y. We denote the set of parent support strategies as $\mathcal{S}$.

2.1 Framework Overview

Our proposed EASRIER is a novel and flexible framework. In this paper, we use MultiAgentESC (Xu et al., 2025a) as an example backbone for empirical evaluation. Figure 1 illustrates the framework design. It first uses an LLM to infer a structured dialogue state $S = (E, C, I)$ from the dialogue context H, where E represents the user's emotional state, C the triggering cause or event, and I the user's underlying intention. We also identify the type of this dialogue turn based on heuristics. We then construct a turn-aware candidate strategy set $\mathcal{S}_s \subseteq \mathcal{S}$ based on the turn type. To capture richer real-world support behavior, we use $\mathcal{S}_s$ as parent strategies and expand them into a larger fine-grained strategy space. In the MCTS-style search, we generate several strategy-conditioned candidate responses during rollouts, evaluate them using LLM-judgment reward, and update the score. Besides, we retrieve similar examples in the training set as few-shot examples, which provide guidance for candidate response generation and to calculate the prior in the score. Finally, we output the top-k branch set as $\mathcal{B}_k = \{(s_i, y_i)\}_{i=1}^{k}$, from which the final response is generated.

2.2 Turn-aware Fine-grained Strategy Space

Previous works have defined eight coarse-grained strategies (Liu et al., 2021), which cannot effectively provide guidance in complicated scenarios. Towards this end, we expand them into a larger set of fine-grained strategies. In addition, we apply a turn-aware policy to filter out obviously mismatched strategies to make the search budget more effective. We first identify the local turn type from the most recent user utterance together with the immediately preceding assistant utterance using heuristics. For example, greetings, thanks, and closing turns are detected by keyword matching, direct questions are detected by question marks and wh-words, question-follow-up turns are detected when a short acknowledgement follows an assistant question, and the remaining turns are assigned to short-content or contentful categories according to utterance length. Based on the resulting turn type, we manually specify a small set of legal parent strategies, a target response length, and a retrieval mode for search. For example, greeting

turns, question-follow-up turns, and contentful support turns require different strategy preferences and different response lengths.

2.3 MCTS-Style Search over Strategy–Response Branches

Given the candidate action set induced by $\mathcal{S}_s$, we perform a lightweight MCTS-style search over the joint strategy–response space. Each branch corresponds to a strategy-conditioned action a. Each rollout selects an action, generates a response, and evaluates the resulting branch.

Selection. Actions are selected using the *Predictor and Upper Confidence bound applied to Trees* (PUCT) score:

$$\mathrm{PUCT}(a) = \frac{Q(a)}{N(a)} + c\,P(a)\frac{\sqrt{N_{\text{tot}}}}{1+N(a)}, \quad (1)$$

where $N(a)$ is the visit count of action a, $Q(a)$ is its accumulated reward, N_{tot} is the total number of rollouts, c controls the exploration–exploitation trade-off, and $P(a)$ is the action prior. Intuitively, $P(a)$ encourages search to focus on actions whose parent strategies are more plausible for the current turn. We compute it in two stages. First, we estimate a prior over parent strategies:

$$P_{\text{parent}}(p) = w_{\text{ret}}P_{\text{ret}}(p)+w_{\text{corp}}P_{\text{corp}}(p)+w_{\text{uni}}P_{\text{uni}}(p),$$

where P_{ret}, P_{corp}, and P_{uni} denote retrieval-based, corpus-based, and uniform parent priors, respectively. Then, for each action a, we use the prior of its parent strategy and normalize it over all candidate actions to obtain $P(a)$. Full details are provided in Appendix A.3.

Simulation and Backup. Given a selected action a, a response y is generated by LLM and evaluated to obtain the reward $r = R(H, S, a, y)$, which is introduced in the next subsection. The statistics $(N(a), Q(a))$ are then updated accordingly. After T rollouts, all sampled branches are ranked by reward, and the top-k candidates form $\mathcal{B}_k$.

2.4 LLM-as-Judge Composite Reward

We define a composite reward produced by LLM-as-a-judge as follows $R(H, S, a, y) \in [0, 10]$, which scores a candidate response using a scalar *overall* score together with auxiliary diagnostic signals. Concretely, an LLM judge is prompted as a rubric-based evaluator to assess 6 perspectives (exploration, insight, action, safety, conciseness, and strategy fidelity) and to output an overall quality score. The rubric aims to align with the goals of ESC, including contextual appropriateness, helpful support, safety, etc. We use a small number of grading examples to guide and stabilize the judgment. Beyond the overall score, the composite reward also incorporates heuristic-based search-time signals, including retrieved-experience similarity, copying penalty, length control, keyword coverage, and response-shape consistency. During search, only the scalar composite reward is used as the rollout return, while the auxiliary signals are retained for interpretable analysis.

Method	D-1↑	D-2↑	B-1↑	B-2↑	B-3↑	F1↑	R-L↑
Zero-shot	6.37	32.02	**17.76**	5.31	2.27	18.20	14.59
Few-shot	6.40	32.22	17.59	5.29	2.27	18.23	14.52
Zero-shot CoT	5.37	24.96	17.41	5.28	2.16	17.12	13.51
Few-shot CoT	5.17	25.91	16.43	4.68	1.97	16.67	12.92
Self-consistency	5.66	30.45	16.16	4.42	1.74	16.53	12.80
Self-Refine	6.13	33.46	16.24	5.08	2.15	17.96	14.31
Mixed-Initiative	6.15	30.39	15.26	4.10	1.61	15.11	11.94
ESCoT	6.58	34.48	15.32	4.53	1.98	16.18	13.02
CogChain	5.79	34.56	16.13	4.54	1.79	16.34	12.99
Cooper	5.71	33.41	16.22	4.89	1.86	16.55	13.34
MAS (Chain)	5.07	29.71	16.59	4.97	2.17	17.26	13.78
MAS (Debate)	5.42	34.14	16.95	5.15	2.16	17.48	13.94
MultiAgentESC	6.47	33.37	17.35	5.35	2.48	17.87	14.13
EASRIER (Ours)	**7.43**	**35.64**	16.68	**6.00**	**2.98**	**18.79**	**15.44**

Table 1: Automatic evaluation on ESConv under the same response-level evaluation setting for all systems.

2.5 Final Candidate Selection

Instead of exhaustively comparing all strategy-conditioned responses, we perform final response selection only over the pre-selected set $\mathcal{B}_k$. The final system response is selected from $\mathcal{B}_k$ through reward-guided reranking together with a lightweight final refinement step.

3 Experiments

3.1 Experimental Setup

Dataset and Metrics. We evaluate on the ESConv benchmark (Liu et al., 2021). To ensure a fair comparison across all systems, we evaluate every system under the same response-level evaluation setting aligned with the released MultiAgentESC output format. Under this setting, the 100-dialogue test slice contains 1210 response instances. For matched ablations, we use a fixed 10-dialogue development slice, which contains 116 response instances under the same setting. Retrieved experience examples and few-shot grading examples used in the judge prompt are constructed only from non-evaluation data. Following MultiAgent-ESC (Xu et al., 2025a), we report two diversity metrics, i.e., Distinct-1 and Distinct-2, and five quality metrics,

Method	D-1↑	D-2↑	B-1↑	B-2↑	B-3↑	F1↑	R-L↑
Top-3 pruned	21.57	**62.83**	**18.23**	**5.94**	**3.10**	**18.93**	**15.27**
Unpruned	**22.91**	61.86	15.17	4.95	2.41	17.47	14.49

Table 2: Ablation studies of our EASRIER on the matched development slice.

i.e., BLEU-1, BLUE-2, BLUE-3, token-level F1 overlap, and ROUGE-L.

Baselines. We compare our proposed EASRIER against a range of baselines including Zero-shot and Few-shot (Brown et al., 2020), Zero-shot CoT (Kojima et al., 2022), Few-shot CoT (Wei et al., 2022), Self-consistency (Wang et al., 2023), Self-Refine (Madaan et al., 2023), Mixed-Initiative (Chen et al., 2023), ESCoT (Zhang et al., 2024), CogChain (Cao et al., 2024), Cooper (Cheng et al., 2024), MAS (Chain) (Xu et al., 2025a), MAS (Debate) (Xu et al., 2025a), and MultiAgentESC (Xu et al., 2025a).

Implementation Details. We use Qwen2.5-32B-Instruct (Qwen et al., 2025). The same backbone is used for dialogue analysis, candidate generation, and judge-style scoring. We use deterministic decoding with a temperature of 0 and a fixed random seed to reduce variance and improve reproducibility. The rollout budget is 12 in the main experiments and 8 in the ablation study. More details are introduced in the appendix.

3.2 Performance Comparison

The compared results of different approaches can be found in Table 1. We have the following observation: ❶ EASRIER outperforms MultiAgentESC and other baselines on six out of seven automatic metrics. This shows that our inference-time search framework is overall more effective than MultiAgentESC. ❷ The clearest gains appear on D-1, D-2, BLEU-2, BLEU-3, F1, and ROUGE-L, indicating that the main advantage of EASRIER comes from better diversity and stronger sequence-level or higher-order overlap with the references, rather than from maximizing unigram overlap alone. ❸ The potential reason for our improvement is that our EASRIER combines turn-aware strategy constraints, fine-grained strategy expansion, and MCTS-style candidate search to compare multiple plausible branches.

3.3 Ablation Study

Our main experiments use the full candidate strategy set constrained by the dialogue turn's type. Here we study a model variant. In particular, *Top-3 pruned search* restricts search to the three highest-priority parent strategy families. In comparison, *Unpruned search* operates on the full candidate parent set. Table 2 shows that Top-3 pruned search outperforms Unpruned search on six out of seven automatic metrics, with the only exception being D-1. These results indicate that a moderate pruning strategy can improve final response quality under our evaluation setting.

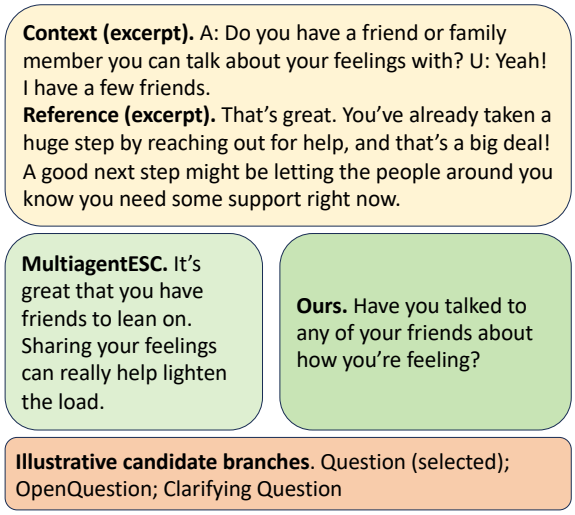

Figure 2: An illustrative example in which the search procedure selects a direct follow-up question, while the baseline remains at generic reassurance.

3.4 Qualitative Results

Figure 2 provides a qualitative example to illustrate the type of behavioral difference induced by search-time candidate selection. In this case, MultiAgentESC remains at generic reassurance, whereas our method selects a more direct follow-up question that is closer to the action-oriented reference response. This pattern is consistent with our motivation for fine-grained strategy search.

4 Conclusion

We propose a novel agentic framework EASRIER for emotional support conversation. It uses the turn-aware strategy constraint, fine-grained strategy space, retrieval-guided priors, MCTS-style shallow candidate search with PUCT score calculation, and a composite reward with LLM-as-a-judge for response generation and selection. Experimental results show that our EASRIER improves the performance over baseline models. Besides, the idea of MCTS-style search to overcome preference bias could also be helpful in broader applications such as multi-agent communication.

5 Limitations

While EASRIER outperforms the baselines, there are still several challenges that need to be addressed in future work. The first challenge to our paper, as well as most papers in this field, is a good evaluation metric that can accurately reflect human users' thoughts. Following previous works, our metrics are automatically calculated based on reference answers and therefore only partially capture real supportive quality. It is expensive to recruit human evaluators, while using LLM-as-a-judge might induce its bias. Second, the running time is longer compared to previous methods because the system performs multiple analysis, retrieval, generation, and judging steps at inference time. Nonetheless, the improved performance already demonstrates that the increased running time is worthwhile. It remains future work to explore an effective and efficient method. Third, the current study is focused on ESConv only, so its generalizability to other supportive dialogue settings is not yet established. Further work is needed to explore ESC agents and generally improve mental health agents.

6 Ethical Considerations

This project concerns emotional support, which is a sensitive application area. The system is intended for research use only and should not be deployed as a replacement for professional mental health care. Responses can be fluent without being clinically appropriate, and reference-based metrics are not sufficient to guarantee user safety. Any practical use would require stronger safety policies, escalation procedures for crisis situations, and careful human oversight.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yaru Cao, Zhuang Chen, Guanqun Bi, Yulin Feng, Min Chen, Fucheng Wan, Minlie Huang, and Hongzhi Yu. 2024. Enhancing emotional support conversation with cognitive chain-of-thought reasoning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 175–187. Springer.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *ACL 2023*.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pages 17853–17861.

Dongjin Kang, Sunghwan Mac 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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 22199–22213.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

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 (ACL-IJCNLP)*, pages 3469–3483.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Navid Madani and Rohini K Srihari. 2025. Steering conversational large language models for long emotional support conversations. In *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, pages 109–123.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Geng Tu, Bingbing Wang, Erik Cambria, Wenjie Li, and Ruifeng Xu. 2025. Supportplay: A multi-agent role-playing system for personalized and sustained multimodal emotional support conversation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2915–2918.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025a. Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yangyang Xu, Zhuoer Zhao, Xiao Sun, and Xun Yang. 2025b. Prompt learning with multiperspective cues for emotional support conversation systems. *IEEE Transactions on Computational Social Systems*.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. *arXiv preprint arXiv:2406.10960*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *Preprint*, arXiv:2304.09582.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. ACL.

A Additional Method Details

A.1 Turn-Type Detection and Strategy Constraints

Our turn-aware policy is rule-based and only uses the local dialogue context. Specifically, we inspect the most recent user utterance together with the immediately preceding assistant utterance. Greeting, thanks, and closing turns are detected by keyword matching; direct questions are detected by question marks and wh-words; question-follow-up turns are detected when a short acknowledgement follows an assistant question; and the remaining turns are assigned to `short_content` or `content` based on utterance length. For each turn type, we manually define a small set of legal parent strategies, a target response length, and a retrieval mode. These constraints are designed to remove clearly mismatched branches before MCTS-style search while preserving reasonable diversity. Table 3 summarizes the turn-type-specific constraints used in our implementation. Each tune type has a preferred target response length, example retrieval mode, max number of rollouts, and a small range of parent strategies.

A.2 Fine-Grained Strategies

We use the eight ESConv strategies as coarse-grained parent strategies, and we expand each parent into a small set of fine-grained child strategies. Table 4 lists the complete parent–child inventory used in our implementation.

A.3 Search Prior Construction

The action prior in Section 2.3 is constructed hierarchically. Let $p(a)$ denote the parent strategy of action a, and let $\mathcal{A}(p)$ denote the set of candidate actions under parent strategy p.

We first compute a parent-level prior:

$$P_{\text{parent}}(p) = w_{\text{ret}} P_{\text{ret}}(p) + w_{\text{corp}} P_{\text{corp}}(p) + w_{\text{uni}} P_{\text{uni}}(p),$$

where: (i) $P_{\text{ret}}(p)$ is a retrieval-based prior estimated from the parent strategies of retrieved experience cases, (ii) $P_{\text{corp}}(p)$ is a corpus prior conditioned on the current turn type, and (iii) $P_{\text{uni}}(p)$ is a uniform smoothing prior over the legal parent strategies.

The turn-aware policy first removes illegal parent strategies for the current turn. Over the remaining parent strategies, we compute $P_{\text{parent}}(p)$ and then distribute the probability mass to their candidate actions. Concretely, for each action a, we assign it the prior of its parent strategy and divide that mass across all actions under the same parent:

$$\tilde{P}(a) = \frac{P_{\text{parent}}(p(a))}{|\mathcal{A}(p(a))|}.$$

Finally, we normalize over all candidate actions:

$$P(a) = \frac{\tilde{P}(a)}{\sum_{a' \in \mathcal{A}} \tilde{P}(a')}.$$

Intuitively, this design combines three signals. The retrieval-based term encourages strategies that are supported by similar retrieved cases, the corpus prior stabilizes the estimate using turn-type-specific statistics, and the uniform prior prevents the search from collapsing too aggressively onto a single strategy family.

B Experiment Setting Details

B.1 Evaluation Setting

All automatic metrics reported in this paper are computed under a shared response-level evaluation setting aligned with the released MultiAgentESC output format. This setting is applied consistently to our system outputs and to all compared baselines. Under this setting, the 100-dialogue test slice contains 1210 response instances, while the matched 10-dialogue development slice used in the ablation studies contains 116 response instances. Our generation is teacher-forced at the dialogue-history level. In particular, for each supporter turn, the model conditions on the gold dialogue history up to that turn, generates one supporter response, and is then evaluated only after alignment and concatenation. This setup keeps the turn-level comparison consistent with the released MultiAgentESC-style output format.

B.2 Data Split

We use the ESConv benchmark (Liu et al., 2021). The final full result is reported on a fixed 100-dialogue test slice. The ablation studies are conducted on a fixed 10-dialogue development slice. Retrieved experience cases and in-context judge examples are constructed only from non-evaluation data. This design prevents evaluation leakage while keeping the comparison setting consistent across all baselines.

B.3 Backbone Model

The backbone model is Qwen2.5-32B-Instruct (Qwen et al., 2025). All experiments

Turn type	Detection rule	Target words	Retrieval mode	Max rollouts	Legal parent strategies
`greeting`	Greeting keywords in the last user turn	8	`last_user`	6	Others, Question
`thanks`	Thanks/appreciation keywords in the last user turn	8	`last_user`	6	Others, Affirmation and Reassurance
`closing`	Closing/sign-off keywords in the last user turn	6	`last_user`	4	Others, Affirmation and Reassurance
`question_followup`	Short acknowledgement after an assistant question	12	`last_pair`	12	Restatement or Paraphrasing, Reflection of feelings, Affirmation and Reassurance, Question, Others, Information; + Providing Suggestions if intent $\in$ {SOLUTION, MIXED} and turn_idx ≥ 1; + Self-disclosure if turn_idx ≥ 2
`short_ack`	Short acknowledgement not classified as question-followup	10	`last_pair`	10	Affirmation and Reassurance, Restatement or Paraphrasing, Reflection of feelings, Others, Question; + Information and Self-disclosure if turn_idx ≥ 2
`question`	Question mark or wh-word in the last user turn	14	`full_context`	18	Information, Question, Restatement or Paraphrasing, Reflection of feelings, Affirmation and Reassurance, Others; + Self-disclosure if turn_idx ≥ 2; + Providing Suggestions if intent $\in$ {SOLUTION, MIXED} and turn_idx ≥ 1
`short_content`	Non-trivial user turn with ≤ 8 tokens	12	`last_pair`	14	Affirmation and Reassurance, Reflection of feelings, Restatement or Paraphrasing, Question, Others, Information; + Self-disclosure if turn_idx ≥ 2; + Providing Suggestions if intent $\in$ {SOLUTION, MIXED} and turn_idx ≥ 1
`content`	All remaining non-trivial turns	18	`full_context`	24	Affirmation and Reassurance, Reflection of feelings, Restatement or Paraphrasing, Question, Providing Suggestions, Others, Information, Self-disclosure; if turn_idx < 2 and intent = SUPPORT, use only Affirmation and Reassurance, Reflection of feelings, Restatement or Paraphrasing, Question, Others, Information

Table 3: Turn-type-specific strategy constraints used by the turn-aware policy.

Parent strategy	Child strategies
Question	Open Question; Clarifying Question
Restatement or Paraphrasing	Summarizing
Reflection of feelings	Reflection - Naming Feelings; Reflection - Link Feelings to Situation
Self-disclosure	Self-disclosure - Similar Experience; Self-disclosure - Shared Feelings
Affirmation and Reassurance	Affirmation - Strengths; Reassurance - Encouragement; Reassurance - Normalizing
Providing Suggestions	Suggestion - Coping Skills; Suggestion - Problem Solving
Information	Information - Resources; Information - Psychoeducation; Information - Data and Facts
Others	Small Talk; Pleasantries - Availability

Table 4: Complete parent–child strategy inventory used for fine-grained search.

use deterministic decoding with temperature 0 to reduce variance and improve reproducibility. The same backbone LLM is used for dialogue analysis, candidate generation, and judge-style scoring.

B.4 Experience Retrieval Setup

For retrieval-based variants, we use a Sentence-Transformer encoder (`all-MiniLM-L6-v2`) to embed query contexts and retrieve the top-10 most similar cases from the non-evaluation pool by cosine similarity. Retrieved cases are used as in-context experience for candidate generation and for calculating the retrieval-guided priors inside the search procedure. The no-RAG ablation disables this retrieval component while keeping the rest of the framework unchanged.

B.5 Search Variants

We compare three search variants in our ablation studies: (1) **Top-3 pruned search**, which keeps only the three highest-priority parent strategy families before the MCTS-style search; (2) **Unpruned search**, which uses all candidate strategies after the turn-aware constraint for the MCTS-style search; (3) **Unpruned search without retrieval** (`copy_guarded`, which disables retrieval under the same development budget.

Table 7 summarizes the main experimental configurations. Except for the difference summarized in this table, other experiment settings all follow

Method	D-1↑	D-2↑	B-1↑	B-2↑	B-3↑	F1↑	R-L↑	Avg. judge↑	Q-ratio↓
Top-3 pruned search	21.57	**62.83**	**18.23**	**5.94**	**3.10**	**18.93**	**15.27**	6.6387	**0.3577**
Unpruned search	**22.91**	61.86	15.17	4.95	2.41	17.47	14.49	**6.7956**	0.6642

Table 5: Extended search-policy ablation study. "Avg. judge" is the average LLM judgement score, and "Q-ratio" means the question-like strategy ratio.

Method	D-1↑	D-2↑	B-1↑	B-2↑	B-3↑	F1↑	R-L↑	Avg. judge↑	Q-ratio↓
Unpruned search (retrieval on)	**22.91**	**61.86**	15.17	4.95	2.41	**17.47**	**14.49**	6.7956	0.6642
Unpruned search (retrieval off)	22.74	61.00	**16.17**	**5.84**	**2.70**	17.16	14.03	**6.8613**	**0.5328**

Table 6: Ablation study of experience retrieval. "Avg. judge" is the average LLM judgement score, and "Q-ratio" means the question-like strategy ratio.

Setting	Rollouts	Retrieval
Unpruned search	12	on
Top-3 pruned search	8	on
Unpruned search (retrieval off)	8	off

Table 7: Overview of the main configurations used in the paper.

the same settings introduced before.

C Additional Ablation Studies

C.1 Extended Search-Policy Ablation

Table 5 presents an extended version of the search-policy ablation from the main text. In addition to the automatic metrics under the shared evaluation setting, we also report raw judge scores and question-like strategy ratios to illustrate the difference between internal selection signals and final task metrics. The extended search-policy ablation highlights an important discrepancy: top-3 pruned search is clearly stronger on the automatic metrics under the shared evaluation setting, but weaker on the average judge score. At the same time, it substantially reduces the question-like strategy ratio. Therefore, our internal judge is useful for search-time guidance, but it should not be treated as the final criterion for model comparison. For final reporting, we prioritize the automatic metrics computed under the shared evaluation setting.

C.2 RAG On/Off Ablation

Table 6 shows that retrieval produces mixed effects. With retrieval enabled, the model performs better on D-1, D-2, F1, and ROUGE-L, suggesting improvements in diversity and sequence-level alignment. In contrast, disabling retrieval yields higher BLEU-1, BLUE-2, BLUE-3, and a higher average judge score, indicating stronger local lexical overlap with the references. Therefore, retrieval should not be viewed as a uniformly positive component. Instead, it changes the trade-off between local lexical matching and broader response quality.

C.3 Metric-Winner Summary

Table 8 summarizes which variant wins which metrics in the matched ablation studies. This summary highlights that neither the search policy nor the retrieval component yields a uniform gain across all automatic metrics. Instead, each design choice changes the balance among diversity, local lexical overlap, and broader sequence-level quality.

D Additional Analysis on Metric–Judge Mismatch

The discrepancy between raw judge scores and automatic metrics is an important observation in our study. Unpruned search receives a higher average judge score than the top-3 pruned search, yet it is clearly weaker on six out of seven automatic metrics. This indicates that the LLM judge captures certain aspects of perceived helpfulness, but does not perfectly match the official evaluation target. A similar phenomenon appears in the ablation study of retrieval. The no-retrieval variant yields a higher average judge score and higher BLEU-1, BLUE-2, and BLUE-3, but retrieval still improves D-1, D-2, F1, and ROUGE-L. This again suggests that different evaluation signals reward different properties of ESC responses.

It might be because the LLM judgment cannot accurately reflect the response quality, or because the LLM judgment is more like a human user's feelings, but the benchmark's automatic evaluation

Comparison	Winner by metric
Search policy	Top-3 pruned search: D-2, B-1, B-2, B-3, F1, R-L Unpruned search: D-1
Retrieval on/off	Retrieval on: D-1, D-2, F1, R-L Retrieval off: B-1, B-2, B-3

Table 8: Summary of metric winners in the matched ablations.

metrics that rely on reference-answer matching are not very good. Future study is needed to explore better evaluation methods.