TEA-Bench: A Systematic Benchmarking of Tool-enhanced Emotional Support Dialogue Agent

Xingyu Sui, Yanyan Zhao*, Yulin Hu, Jiahe Guo, Weixiang Zhao, Bing Qin
Harbin Institute of Technology
{xysui, wxzhao, yyzhao}@ir.hit.edu.cn

Abstract

Emotional Support Conversation requires not only affective expression but also grounded instrumental support to provide trustworthy guidance. However, existing ESC systems and benchmarks largely focus on affective support in text-only settings, overlooking how external tools can enable factual grounding and reduce hallucination in multi-turn emotional support. We introduce **TEA-Bench**, the first interactive benchmark for evaluating tool-augmented agents in ESC, featuring realistic emotional scenarios, an MCP-style tool environment, and process-level metrics that jointly assess the quality and factual grounding of emotional support. Experiments on nine LLMs show that tool augmentation generally improves emotional support quality and reduces hallucination, but the gains are strongly capacity-dependent: stronger models use tools more selectively and effectively, while weaker models benefit only marginally. We further release **TEA-Dialog**, a dataset of tool-enhanced ESC dialogues, and find that supervised fine-tuning improves in-distribution support but generalizes poorly. Our results underscore the importance of tool use in building reliable emotional support agents. [1]

1 Introduction

In modern society, people increasingly experience emotional stress due to mounting pressures from work and daily life. Thus, the demand for Emotional Support Conversation (ESC) (Liu et al., 2021) has grown substantially, as they offer psychological relief, reassurance and guidance during moments of distress (Langford et al., 1997; Greene and Burleson, 2003; Heaney and Israel, 2008).

Social support theory distinguishes two complementary types of support in ESC: **affective support**, which conveys empathy and care, and **instrumental support**, which offers concrete guidance and actionable assistance (Cutrona and Russell, 1990; Semmer et al., 2008). While affective support can be expressed through empathetic language alone, instrumental support requires accurate grounding in real-world situational information. Without such grounding, advice may become generic or factually incorrect, undermining user trust (Klyver et al., 2017; Shaikh et al., 2024).

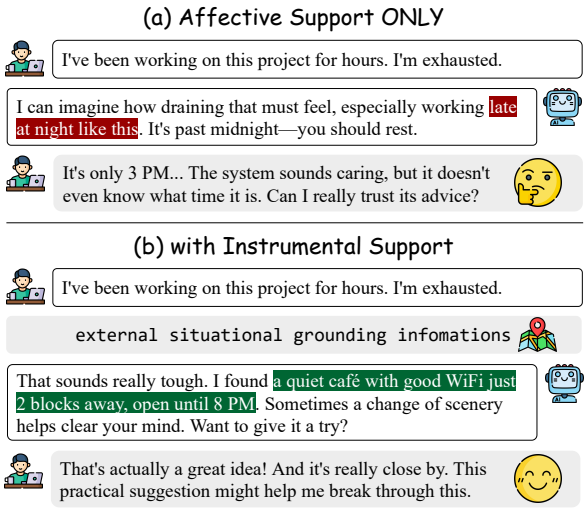

Figure 1: Comparison of affective-only support vs. instrumental support. (a) Affective-only response hallucinates details, undermining trust. (b) Instrumental support response provides verified, actionable suggestions.

However, existing ESC systems predominantly focus on affective support alone. Despite advances in linguistic fluency and empathetic expression, recent models operate largely in text-only settings with limited access to external contextual information (Zhao et al., 2023a; Chen et al., 2023; Farhat, 2024). As illustrated in Figure 1(a), affective-only responses often contain unwarranted hallucinated contextual assumptions that damage credibility despite sounding caring. In contrast, Figure 1(b) shows how responses grounded with instrumental support leverage verified external information

* Corresponding author
[1] Our code and data can be found in https://github.com/XingYuSSS/TEA-Bench.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46390–46416
July 2-7, 2026

to provide trustworthy, actionable suggestions that complement emotional validation.

Recent advances in tool-augmented large language models (LLMs), provide a promising pathway toward addressing this gap by allowing models to dynamically acquire external situational knowledge during interaction (Mialon et al., 2023; Qin et al., 2024). Nevertheless, existing evaluation frameworks remain inadequate. ESC benchmarks predominantly assess text-only empathy (Zhao et al., 2024), while evaluations of tool-augmented systems mainly focus on task-oriented performance (Zhang et al., 2024; Li et al., 2023; Trivedi et al., 2024). As a result, the role of tools in enabling grounded, empathetic, and multi-turn emotional support remains poorly understood.

To address this gap, we introduce `TEA-Bench` (Tool-enhanced Emotional Support Dialogue Agent Benchmark), the first interactive benchmark for evaluating tool-augmented agents in ESC. `TEA-Bench` investigates how access to external tools facilitates grounded, empathetic, and context-aware support in multi-turn interactions. The benchmark comprises fine-grained emotional support scenarios adapted from ExTES (Zheng et al., 2024), a realistic MCP-based tool environment, and a simulated user that provides structured feedback, including reactions to hallucinated or inappropriate responses. Crucially, `TEA-Bench` enables process-level analysis through complementary metrics that assess both ESC quality and factual grounding, capturing model behavior across conversational turns.

Using `TEA-Bench`, we evaluate nine contemporary LLMs, including four closed-source and five open-source models. We identify capacity-dependent patterns in tool-enhanced emotional support: tool augmentation improves ESC performance and reduces hallucination, but the gains depend on models ability to invoke and integrate tools effectively. Stronger models leverage precise, selective tool usage, while mid-capability models rely on more frequent calls to achieve comparable grounding; weaker models struggle to benefit and show limited empathy gains. Further analysis reveals a positive relationship between tool usage and hallucination mitigation, with substantial efficiency differences across model scales.

We also curate `TEA-Dialog`, a high-quality dataset of grounded, tool-enhanced dialogues derived from `TEA-Bench` interactions. While supervised fine-tuning on this dataset improves empathetic performance in familiar scenarios, it exhibits limited generalization and may increase hallucination under distribution shift, highlighting both the promise and challenges of training reliable tool-augmented emotional support agents.

This work makes three main contributions: (1) we articulate the need for grounded instrumental support in emotional support conversations and highlight the lack of evaluation frameworks that capture how external tools enable context-aware, trustworthy empathy in multi-turn ESC; (2) we introduce `TEA-Bench`, the first interactive benchmark designed to evaluate tool-augmented emotional support agents, featuring realistic scenarios, an MCP-style tool environment, and process-level metrics for empathy and hallucination; and (3) through extensive experiments on nine contemporary LLMs, we uncover capacity-dependent patterns in tool utilization and hallucination mitigation, and release `TEA-Dialog`, a high-quality dataset of grounded, tool-enhanced ESC dialogues to support future research and training.

2 Related Works

Emotional Support Conversation. Emotional Support Conversations (ESC) (Liu et al., 2021) study interactions between a seeker experiencing emotional distress and a supporter aiming to alleviate emotional intensity through appropriate conversational strategies. Early work explored task-specific architectures and modeling techniques, such as hierarchical graph networks (Peng et al., 2022; Zhao et al., 2022), commonsense knowledge integration (Tu et al., 2022), and joint emotion-semantic modeling (Zhao et al., 2023b). With the emergence of large language models (LLMs), recent approaches leverage pretrained models via supervised fine-tuning on curated ESC datasets, improving multi-turn coherence and empathetic quality without architectural modification (Liu et al., 2023; Chen et al., 2023; Qiu et al., 2023). Beyond supervised learning, Zhao et al. (2025) propose a Chain of Strategy Optimization (CSO) framework that uses Monte Carlo Tree Search to generate high-quality preference data and applies preference optimization, further enhancing ESC performance. However, these approaches primarily operate in text-only settings and do not consider the role of external grounding or tool use in emotional support.

Figure 2: Example illustrating a tool-augmented ESC under our benchmark setting. The assistant dynamically invokes external tools to gather contextual information, enabling it to offer emotionally resonant and actionable support rather than generic reassurance.

Benchmarks for Tool-augmented LLM. Tool-augmented large language models enable interaction with external tools and real-world information sources, substantially extending their capabilities (Mialon et al., 2023; Qin et al., 2024). Prior benchmarks mainly evaluate task-oriented tool use, focusing on accurate invocation, execution correctness, and task completion (Zhang et al., 2024; Li et al., 2023; Trivedi et al., 2024; Liu et al., 2025; Han et al., 2025). However, these evaluations are largely designed for instruction-following or problem-solving scenarios and pay limited attention to human-centered interaction. How tool grounding should adapt to users emotional states and situational constraints in multi-turn supportive dialogue remains underexplored, motivating dedicated evaluation frameworks for emotional support settings.

3 TEA-Bench

3.1 Task Definition

Figure 2 provides an illustrative example of a tool-augmented emotional support conversation under our benchmark. `TEA-Bench` evaluates agents in emotional support conversations between a user and an agent. At each turn, the agent first decides whether external information is needed and may optionally invoke tools to retrieve situational or factual information. Any retrieved tool outputs are incorporated into the agents context. Based on the dialogue history and this contextual information, the agent then generates an appropriate response. Tool use serves as auxiliary support for grounding rather than completing an explicit task.

Asymmetric Information Access. The agent observes the full interaction history including all tool invocations and outputs. In contrast, users and evaluators observe only the natural language utterances, mirroring real-world scenarios where internal information retrieval is hidden from users.

Evaluation Objectives. Agents are evaluated on: (1) **empathetic support quality**, consistently providing emotionally appropriate and helpful responses; and (2) **factual grounding**, ensuring concrete claims are traceable to user-provided information or tool observations, avoiding hallucinated content. Unlike task-oriented dialogues, no explicit terminal goal is defined, reflecting the open-ended nature of emotional support.

3.2 Scenario Construction

We construct 81 grounded emotional support scenarios, referred to as `TEA-Scenarios`, from the ExTES dataset (Zheng et al., 2024) through a four-stage pipeline shown in Figure 3. Our goal is to enrich each scenario with retrievable spatiotemporal context, enabling agents to provide factually grounded support while preserving the authenticity of emotional support interactions.

Scenario Filtering. We retain ExTES scenarios with situation descriptions exceeding 30 words. This threshold ensures sufficient contextual richness for both empathetic reasoning and meaningful tool usage, filtering out underspecified cases where grounding opportunities would be limited.

Latent Context Generation. For each retained scenario, we use an LLM to infer implicit situational attributes that are relevant but not explicitly stated in emotional support contexts, including the users local time, approximate city-level location, and the type of place the user is in, such as a workplace or public environment. These attributes are generated by conditioning on the original scenario description and serve as latent contextual variables

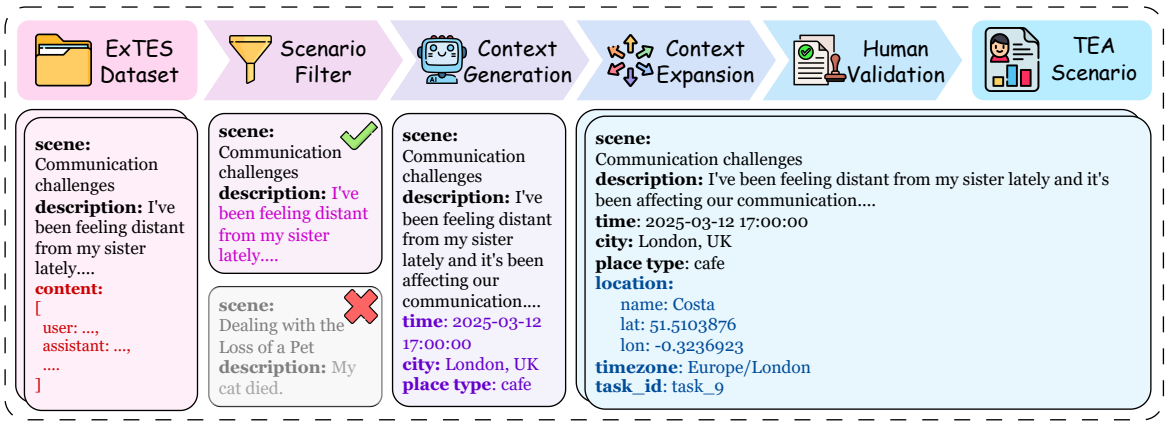

Figure 3: Overview of TEA-Scenario construction pipeline. We filter emotionally rich scenarios, generate latent spatiotemporal attributes via LLM, ground them through map-based APIs, and validate through human review.

that agents can retrieve through tools. The generation prompt is provided in Appendix G.1.

Context Grounding. The inferred location information is subsequently grounded using map-based API queries to obtain concrete geographic data, including precise coordinates, verified place names, and corresponding time zones. Each scenario is assigned a unique identifier to ensure consistent and reliable information retrieval.

Human Validation. All constructed scenarios undergo manual review by the authors to ensure quality and realism. We filter out cases with implausible location-situation pairings or internally inconsistent temporal information. Only scenarios that are coherent, contextually realistic, emotionally plausible, and suitable for emotional support interactions are retained in the final benchmark.

3.3 Tool Environment

To support grounded instrumental support, TEA-Bench provides a diverse tool environment that allows agents to retrieve contextual information beyond the dialogue history.

Tool Design. The environment includes 31 tools across seven categories: Reddit, Map, Utils, Weather, News, Wikipedia, and Music. Appendix A details the complete tool list and their data sources. Each category serves a distinct grounding function. Reddit tools retrieve shared experiences from online communities. Map and Weather tools provide spatial and environmental information. News tools offer situational awareness. Wikipedia tools supply encyclopedic knowledge. Music tools support affective content recommendation. All tools expose real-world data sources via Model Context Protocol (MCP).

Scenario-Aware Execution. Time-sensitive tools (Reddit, Weather, News) execute in a scenario-aware manner to ensure reproducibility while preserving temporal realism. Tool calls implicitly condition on the timestamp of the current TEA-Scenario rather than the system clock. They return results corresponding to that specific moment. This mechanism provides a consistent external environment across agents while maintaining realistic time-dependent dynamics.

Context Retrieval Utilities. The Utils category retrieves scenario-specific contextual information without calling external APIs. These utilities directly access the latent attributes from Section 3.2, including the user's local time, geographic coordinates, and current place. This simulates device-level information access that agents would have in real deployment. Agents can obtain grounded context through Utils before invoking downstream tools like Map or Weather.

Open-Ended Tool Usage. Tool use is entirely optional and unsupervised. Agents may invoke any subset of tools at any turn. They must infer from conversational context whether, when, and which tools to use. This reflects realistic deployment settings where appropriate instrumental support must be judged by the agent.

3.4 Interactive Evaluation Framework

TEA-Bench evaluates agents through interactive multi-turn dialogues with a simulated user. As shown in Figure 4(a), agents may invoke tools at any turn to retrieve contextual information. The agent is follows a fixed system prompt that encourages proactive tool usage and concise responses. Appendix G.2 provides the full prompt.

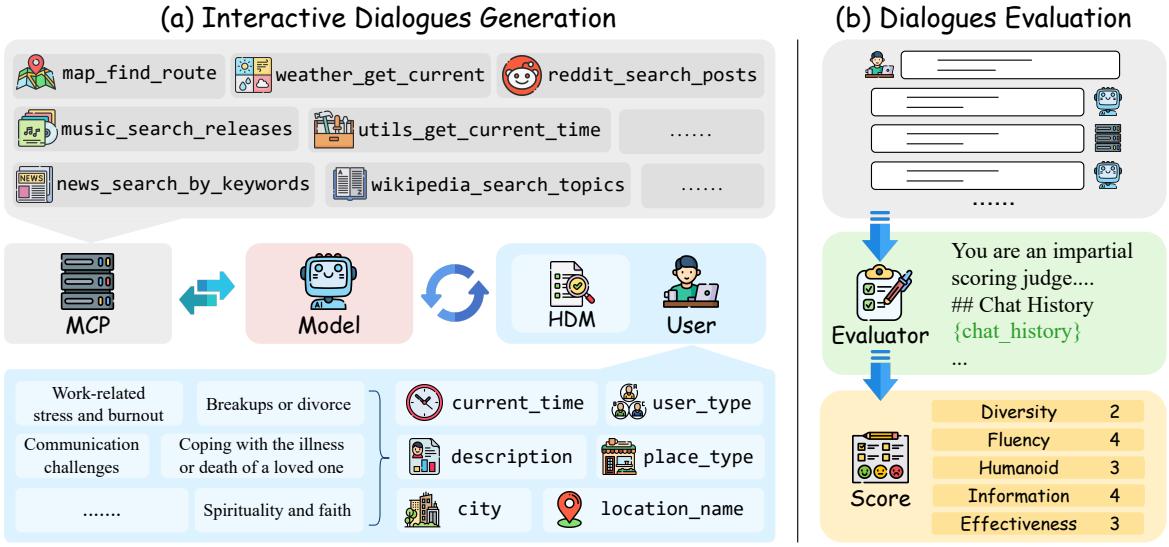

Figure 4: Overview of TEA-Bench. The evaluated agent engages in multi-turn emotional support dialogues with a simulated user and may invoke external tools via the Model Context Protocol (MCP). A Hallucination Detection Module verifies factual grounding in agent responses based on dialogue history and tool observations. Complete dialogues are evaluated using the TEA score and factuality metrics.

Tool Interaction. Tool usage is entirely controlled by the agent's native function-calling capability. All MCP tools are exposed as callable functions without prompt-level supervision. At each turn, the agent may generate a response or invoke tools. Tool observations are appended to the agent's context. The agent may continue calling tools or produce a final response.

Hallucination Detection Module (HDM). We introduce a Hallucination Detection Module (HDM) to identify whether factual content in agent responses is grounded in available context. A response is grounded if all factual entities (including locations, times, events, external conditions) can be traced to user-provided information or tool observations in the dialogue history.

Two types of content are excluded from HDM detection. First, commonsense elaborations that do not introduce new factual entities. Second, general emotional support or lifestyle suggestions that do not rely on external world states. The HDM employs a prompted language model that verifies entity provenance against dialogue and tool context. Appendix G.3 provides the detection prompt.

User Simulation. User behavior is simulated using a language model conditioned on scenario attributes from Section 3.2, including local time, geographic context, user type, and the emotional situation description. The simulator generates user utterances based on dialogue history and expresses doubt when hallucination is detected using a hallucination-aware prompt.

We define two user types. *Action-oriented* users regulate emotions through actions and may not immediately articulate causes, while *Emotion-oriented* users require emotional validation before accepting practical suggestions. The simulator may terminate the dialogue when it deems the interaction complete. Appendix G.4 provides user simulation prompts and type definitions.

Episode Generation. Each episode begins with an initial user utterance. Agent and user then alternate turns. Within each agent turn, the model may invoke tools multiple times before responding. Dialogue length is capped at 15 turns unless the simulator terminates earlier.

3.5 Evaluation Metrics

TEA-Bench evaluates agents at the dialogue level, jointly considering empathetic support quality and factual grounding. As shown in Figure 4(b), we use a language model as an automatic evaluator to score complete dialogue episodes.

TEA-Scores. We assess empathetic support quality using five key dimensions. Four are directly adapted from ESC-Eval (Zhao et al., 2024): **Diversity**, **Fluency**, **Humanoid**, and **Information**. We introduce the fifth dimension, **Effectiveness**, which measures whether the agent's suggestions are accepted and integrated by the user.

The first three dimensions assess dialogue quality and the agent's affective support, while the lat-

Model	TEA-Scores					AVG.(TEA) ↑	Factuality		
	Div. ↑	Flu. ↑	Hum. ↑	Info. ↑	Eff. ↑		Fact. ↑	Halluc. ↓	**Halluc. Rate ↓**
without tool									
GPT-4o-mini	65.90	86.88	91.36	**63.27**	75.93	76.67	**82.38**	21.60	24.95
GPT-4.1-nano	69.14	**90.74**	**95.68**	**63.27**	**81.94**	**80.15**	62.21	11.35	14.51
Gemini-2.5-flash	58.33	80.09	78.55	57.10	59.57	66.73	75.71	54.59	64.92
Qwen-plus	66.05	83.49	85.03	60.03	68.83	72.69	75.98	57.26	68.81
Qwen3-235B-a22B	67.13	84.26	83.64	59.26	65.12	71.88	79.48	61.05	71.21
Qwen3-next-80B-a3B	**69.60**	86.42	88.89	60.34	80.09	77.07	64.44	41.85	62.49
Qwen3-32B	64.81	85.49	89.04	59.10	73.77	74.44	69.23	37.58	50.60
Qwen3-14B	64.66	87.35	92.44	58.95	80.71	76.82	59.81	15.68	21.76
Qwen3-8B	65.90	87.96	93.98	57.56	81.79	77.44	52.28	**8.78**	**11.41**
with tool									
GPT-4o-mini	70.22	**91.05**	95.52	**68.06**	80.71	81.11	77.61	14.72	18.44
GPT-4.1-nano	70.22	90.43	**96.91**	64.2	**84.26**	**81.2**	57.6	8.64	11.69
Gemini-2.5-flash	66.05	89.04	91.51	61.57	79.48	77.53	65.74	17.25	21.01
Qwen-plus	68.83	88.89	90.12	63.89	78.7	78.09	**83.24**	30.05	34.89
Qwen3-235B-a22B	68.98	89.51	91.67	65.43	81.02	79.32	78.42	25.5	31.44
Qwen3-next-80B-a3B	**71.76**	89.97	91.2	61.57	79.17	78.73	69.44	40.85	57.57
Qwen3-32B	67.59	89.51	94.29	59.57	84.1	79.01	68.21	24.13	34.03
Qwen3-14B	63.58	87.35	94.75	56.02	80.09	76.36	53.89	8.57	12.45
Qwen3-8B	65.9	87.81	95.37	58.18	81.64	77.78	51.95	**5.41**	**7.16**

Table 1: Evaluation on TEA-Bench. For the *with tool* setting, blue cells indicate improvement compared to *without tool*, while red cells indicate performance drop. The intensity of the color reflects the magnitude of the change. **Bold** and underlined numbers denote the best and second-best results, respectively.

ter two evaluate instrumental support through the quality and relevance of recommendation. Each dimension receives an integer score from 0 to 4 based on the full dialogue. Raw scores are normalized to 0-100 and averaged to produce the TEA score. Appendix B.1 provides detailed rubrics and Appendix G.5 shows the evaluation prompt.

Factuality Metrics. We quantify factual grounding using three dialogue-level statistics: (i) **factual content ratio**, the proportion of agent responses containing factual claims; (ii) **hallucination ratio**, the proportion of responses containing hallucinated content; (iii) **hallucination rate**, the ratio of hallucinated content to all factual content.

A factual claim is considered hallucinated if it cannot be traced to user-provided information or tool observations (Section 3.4). All metrics are averaged across dialogues, ensuring that each interaction contributes equally. Together, these metrics characterize how frequently agents include factual information and how reliably it is grounded. Appendix B.2 provides detailed definitions.

4 Experiments

4.1 Experimental Setup

We evaluate various LLMs under TEA-bench, including four closed-source models *GPT4o-mini* (Hurst et al., 2024), *GPT-4.1-nano* (OpenAI, 2025), *Gemini-2.5-flash* (Comanici et al., 2025) and *Qwen-plus* (Yang et al., 2025), as well as five open-source models *Qwen3-235B-A22B-Instruct-2507* (Yang et al., 2025), *Qwen3-Next-80B-A3B-Instruct* (Yang et al., 2025), *Qwen3-32B* (Yang et al., 2025), *Qwen3-14B* (Yang et al., 2025) and *Qwen3-8B* (Yang et al., 2025).

4.2 RQ1: How Well Do Current Models Perform in Tool-enhanced ESC?

This research question examines whether tool augmentation improves ESC performance. We evaluate nine representative models under two settings: *without tool* and *with tool*. The main results are reported in Table 1. The reliability of these automatic metrics is validated through human evaluation, as detailed in Appendix C.

Overall Impact of Tool Augmentation. Overall, enabling tool use improves TEA-Scores for most evaluated models. Models with strong reasoning and tool-calling capabilities exhibit substantial gains, indicating that tools provide useful contextual grounding for empathetic responses. In contrast, models with weaker tool usage ability may experience limited improvement or slight degradation, suggesting that effective tool integration is a prerequisite for performance gains.

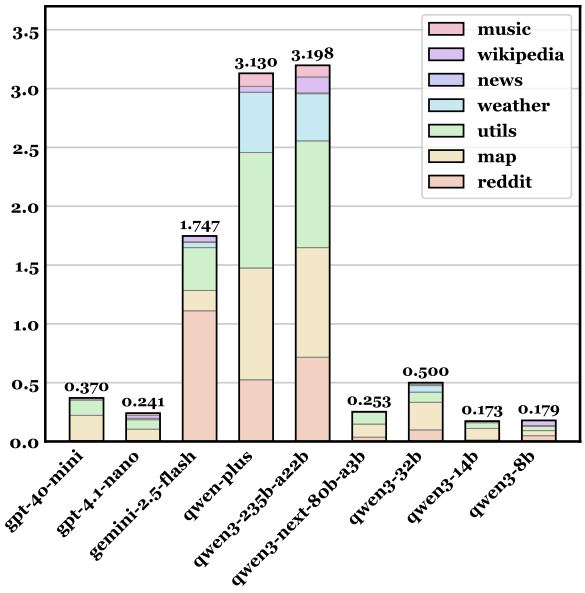

Figure 5: Average number of tool calls per dialogue across different models on TEA-Bench.

Hallucination and Factual Grounding. Across all models, tool augmentation consistently reduces hallucination-related metrics. Both the proportion of hallucinated factual content and the hallucination rate decrease under the *with tool* setting, demonstrating that tools reliably enhance factual grounding even when empathy gains are modest.

Results Across User Types. We further report results under different user types in Appendix D.1. Tool augmentation yields particularly strong improvements for action-oriented users, while performance for emotion-oriented users improves for most models but may decline slightly for weaker ones. Importantly, hallucination rates are consistently reduced for both user types.

These results demonstrate that tool augmentation can effectively enhance overall ESC performance and factual grounding, motivating further analysis of tool usage behaviors in RQ2.

4.3 RQ2: How Do Tools Improve ESC?

This research question investigates how tools contribute to improved ESC performance and hallucination mitigation. We analyze tool usage behaviors across models, their relationship with hallucination reduction, and the characteristics of high-quality tool-enhanced dialogues.

Tool Usage Frequency across Models. Figure 5 compares the average number of tool calls per dialogue across different models. Stronger models achieve substantial performance gains

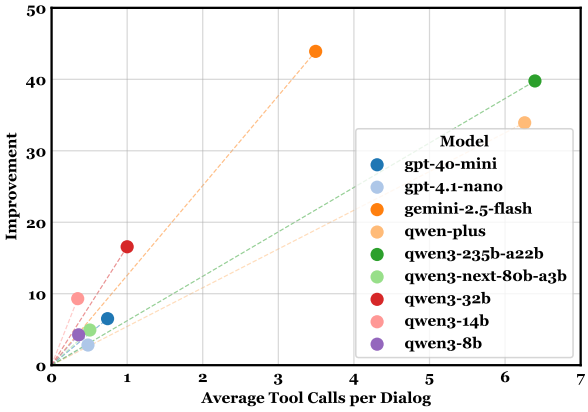

Figure 6: Relationship between tool usage frequency and hallucination rate reduction across models. Dashed lines to the origin indicate per-call efficiency, with steeper slopes corresponding to higher efficiency.

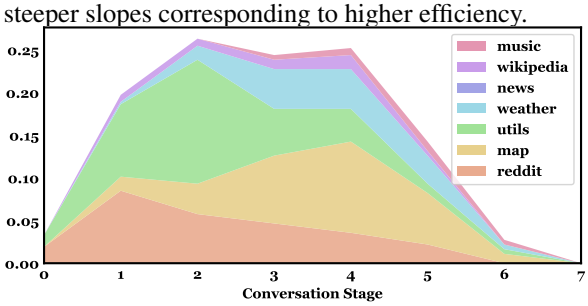

Figure 7: Average tool usage distribution across normalized dialogue stages in TEA-Dialog. Different colors represent different tool categories.

with relatively fewer tool calls, indicating more precise and effective tool usage. In contrast, mid-capability models rely on more frequent tool invocations to obtain comparable improvements, while weaker models invoke tools infrequently and show limited benefits, reflecting difficulties in effectively leveraging external tools. We further report results under different user types in Appendix D.2.

Tool Usage and Hallucination Reduction. We further examine the relationship between tool usage frequency and hallucination mitigation. As shown in Figure 6, the reduction in hallucination rate generally increases with the number of tool calls, revealing a clear positive correlation between tool usage and hallucination elimination. However, models vary substantially in efficiency: stronger models achieve larger hallucination reductions with fewer tool calls, whereas mid-capability models require more frequent interactions to obtain similar gains. Smaller models exhibit limited absolute hallucination reduction despite occasional high per-call efficiency.

TEA-Dialog: High-quality Tool-enhanced Dialogue Dataset. To better understand effective

Model	TEA-Bench					AVG.(TEA) ↑	Hallucination		
	Div. ↑	Flu. ↑	Hum. ↑	Info. ↑	Eff. ↑		Fact. ↑	Halluc. ↓	**Halluc. Rate ↓**
Qwen3-8B									
Base	65.9	87.81	95.37	58.18	81.64	77.78	51.95	5.41	7.16
TEA	69.91	87.19	92.59	60.34	82.56	78.52	68.65	6.86	9.73
TEA-ID	71.67	88.54	94.17	62.71	82.5	79.92	70.38	3.73	4.61
TEA-OOD	64.88	83.33	88.1	53.57	82.74	74.52	63.71	15.8	24.39
Qwen3-14B									
Base	63.58	87.35	94.75	56.02	80.09	76.36	53.89	8.57	12.45
TEA	66.98	86.88	93.36	60.34	82.41	77.99	66.32	9.12	12.2
TEA-ID	67.92	86.88	94.17	60.42	82.29	78.34	68.09	7.4	9.86
TEA-OOD	64.29	86.9	91.07	60.12	82.74	77.02	61.25	14.03	18.89

Table 2: Evaluation of TEA-trained models on in-domain (ID) and out-of-domain (OOD) scenarios. blue cells indicate improvement compared to Base, while red cells indicate performance drop. The intensity of the color reflects the magnitude of the change.

tool usage patterns, we collect TEA-Dialog, a dataset of 365 high-quality dialogues. Dialogues are selected from multiple models based on high TEA-Scores (above 80) and absence of detected hallucinations, with additional human filtering to ensure reliability.

Figure 7 shows the average tool usage across normalized dialogue stages. Early stages primarily involve utility and contextual tools to establish situational grounding, followed by environment-aware tools such as maps and weather. In later stages feature personalized tools, including music and news, reflecting a structured transition from context acquisition to tailored emotional support. Additional analyses on tool usage patterns under different user types and other dialogue attributes are provided in Appendix E.

4.4 RQ3: How Can Tool-enhanced ESC Be Further Improved?

We investigate whether supervised fine-tuning (SFT) on TEA-Dialog can further improve ESC performance. Using dialogues from the first 60 scenarios, we fine-tune Qwen3-8B and Qwen3-14B, yielding Qwen3-8B-TEA and Qwen3-14B-TEA. Models are evaluated on all 81 scenarios, divided into **In-Domain (ID)** scenarios seen during training and **Out-Of-Domain (OOD)** scenarios with unseen descriptions. Training details are in Appendix F.

Overall Performance. Both TEA-trained models achieve consistent improvements over their base counterparts across TEA-Scores, particularly *Information* and *Effectiveness*. These gains are accompanied by a substantial increase in factual content, indicating that SFT encourages models to produce more grounded responses.

Generalization under Distribution Shift. Performance on ID scenarios is significantly higher than on OOD scenarios for both models, suggesting limited generalization when training on a small set of high-quality scenarios. Notably, Qwen3-14B exhibits a smaller gap between ID and OOD than Qwen3-8B, indicating that larger models generalize better at the scenario level.

Hallucination Trade-offs. While hallucination rates decrease on ID scenarios after fine-tuning, they increase markedly on OOD scenarios. This trend suggests that although SFT improves factual grounding in familiar contexts, it may also amplify hallucination risks under distribution shift.

These results indicate that naïve SFT on limited high-quality data is insufficient for robustly improving tool-enhanced ESC, motivating the need for more effective training or alignment strategies.

5 Conclusion

We emphasize ESC with instrumental support to ensure trustworthiness and actionable guidance. Existing systems focus on affective expression in text-only settings, leaving factual grounding and hallucinations underexplored. We introduce TEA-Bench, a benchmark for tool-augmented ESC agents, and TEA-Dialog, a dataset of tool-enhanced dialogues. Experiments across nine LLMs show tool augmentation improves support quality and reduces hallucination, with gains dependent on model capacity. Fine-tuning on

TEA-Dialog boosts in-distribution performance but generalizes poorly, highlighting challenges of reliable, tool-augmented ESC agents.

Limitations

Simulated User. TEA-Bench relies on a simulated user to enable controlled, reproducible, and scalable evaluation. While this allows for fine-grained and process-level assessment, it cannot fully capture the diversity, spontaneity, and unpredictability of real human behavior in ESC.

Interaction Horizon. Our evaluation focuses on short to medium-length ESC interactions. Long-term dynamics such as sustained emotional support, trust formation, and user adaptation across extended conversations are not explicitly modeled and remain an important direction for future work.

Generalization of Training Data. Although fine-tuning on TEA-Dialog improves performance in familiar scenarios, we observe limited generalization under distribution shift, and in some cases increased hallucination. This suggests that supervised fine-tuning alone may be insufficient for training robust and broadly generalizable tool-augmented emotional support agents.

Long-term Memory. TEA-Bench focuses on immediate situational grounding via tools rather than long-term conversational memory. Recent benchmarks (Maharana et al., 2024; Guo et al., 2026; Jiang et al., 2025; Hu et al., 2026) have systematically evaluated memory recall and retention across extended dialogues, which remains unexplored in our work.

Ethical Statement

We are committed to publicly releasing all data upon acceptance of the paper. We are fully aware of the potential biases associated with LLM-as-Judge. To mitigate these effects, we incorporated human expert assessments. However, due to cost considerations, the scale of human evaluation remains limited at this stage. We note that this constraint is common in current conversational AI research that relies on LLMs.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions. This work was supported by the National Natural Science Foundation of China (NSFC) via grant 62441614 and 62576125.

References

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Carolyn E Cutrona and Daniel W Russell. 1990. Type of social support and specific stress: Toward a theory of optimal matching.

Faiza Farhat. 2024. Chatgpt as a complementary mental health resource: a boon or a bane. *Annals of Biomedical Engineering*, 52(5):1111–1114.

John O Greene and Brant R Burleson. 2003. *Handbook of communication and social interaction skills*. Routledge.

Jiahe Guo, Xiangran Guo, Yulin Hu, Zimo Long, Xingyu Sui, Xuda Zhi, Yongbo Huang, Hao He, Weixiang Zhao, Yanyan Zhao, and 1 others. 2026. When personalization legitimizes risks: Uncovering safety vulnerabilities in personalized dialogue agents. *arXiv preprint arXiv:2601.17887*.

Han Han, Tong Zhu, Xiang Zhang, MengSong Wu, Xiong Hao, and Wenliang Chen. 2025. NesTools: A dataset for evaluating nested tool learning abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9824–9844, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health behavior and health education: Theory, research, and practice*, 4(1):189–210.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yulin Hu, Zimo Long, Jiahe Guo, Xingyu Sui, Xing Fu, Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2026. Op-bench: Benchmarking over-personalization for memory-augmented personalized conversational agents. *arXiv preprint arXiv:2601.13722*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*.

Kim Klyver, Benson Honig, and Paul Steffens. 2017. Social support timing and persistence in nascent entrepreneurship: exploring when instrumental and emotional support is most effective. *Small Business Economics*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Catherine Penny Hinson Langford, Juanita Bowsher, Joseph P Maloney, and Patricia P Lillis. 1997. Social support: a conceptual analysis. *Journal of advanced nursing*, 25(1):95–100.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. API-bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116, Singapore. Association for Computational Linguistics.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

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.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2025. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *Preprint*, arXiv:2302.07842.

OpenAI. 2025. Introducing GPT-4.1 in the API. Blog post. Accessed: 2025-12-30.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, and 24 others. 2024. Tool learning with foundation models. *ACM Comput. Surv.*, 57(4).

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Norbert K Semmer, Achim Elfering, Nicola Jacobshagen, Tanja Perrot, Terry A Beehr, and Norbert Boos. 2008. The emotional meaning of instrumental social support. *International journal of stress management*, 15(3):235.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. 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 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076,

Bangkok, Thailand. Association for Computational Linguistics.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, Tetsuya Sakai, Tian Feng, and Hayato Yamana. 2024. ToolBeHonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11388–11422, Miami, Florida, USA. Association for Computational Linguistics.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, and 1 others. 2024. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810.

Weixiang Zhao, Xingyu Sui, Xinyang Han, Yang Deng, Yulin Hu, Jiahe Guo, Libo Qin, Qianyun Du, Shijin Wang, Yanyan Zhao, and 1 others. 2025. Chain of strategy optimization makes large language models better emotional supporter. *arXiv preprint arXiv:2503.05362*.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *IJCAI*, pages 4524–4530.

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

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

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

A Tool Categories and Data Sources

Table 3 summarizes the concrete contents provided by each tool category. News-related tools are backed by the GDELT Project, providing access to global news events. Weather information is obtained from Meteostat, while music-related queries rely on MusicBrainz metadata. Map information is supported by OpenStreetMap, and community discussions are retrieved from Reddit. Encyclopedic knowledge is accessed via Wikipedia. Within the `Utils` category, `utils_get_current_time` and `utils_get_user_location` read predefined scenario attributes, while `utils_fetch_webpage_content` extracts plain text from a specified webpage using `trafilatura`.

B Details of Metrics

B.1 TEA Metrics.

The explanations of each metric are as follows:

Diversity (Div.) Focusing on the diversity of expression forms and the richness of content in dialogue.

Fluency (Flu.) Not only focus on the logical coherence of the context in dialogues but also pay attention to the fluency of expression in a given conversation.

Humanoid (Hum.) Focus on the differences between emotional assistants and humans.

Information (Info.) Focusing on Evaluating the Reasonableness and Quantity of Recommendations Provided by Emotion Assistants.

Effectiveness (Eff.) Focusing on whether the supporters suggestions are practically accepted and meaningfully integrated by the help-seeker, as reflected in their subsequent responses or emotional shifts.

Evaluation rules are listed in Table 4.

B.2 Factuality and Hallucination Metrics.

Let a dialogue d consist of a sequence of model responses $\{r_1, r_2, \ldots, r_{T_d}\}$. For each response, we annotate whether it contains factual content and whether such content is hallucinated according to the grounding criteria described in Section 3.4.

For each dialogue d, we define:

Wikipedia	wikipedia_search_topics
	wikipedia_get_summary
	wikipedia_get_section
	wikipedia_get_available_sections
	wikipedia_get_full_content
Map	map_find_route
	map_find_nearby_places
	map_calculate_reachable_area
	map_get_location_info
Weather	weather_get_current
	weather_get_forecast
Utils	utils_get_current_time
	utils_get_user_location
	utils_fetch_webpage_content
News	news_search_by_keywords
	news_get_related_themes
	news_search_by_theme
	news_search_by_location
Music	music_search_artists
	music_search_releases
	music_search_recordings
	music_search_releases_by_year
	music_get_artist_details
	music_get_release_details
	music_get_recording_details
Reddit	reddit_search_posts
	reddit_search_subreddit
	reddit_get_subreddit_posts
	reddit_get_post_comments
	reddit_get_subreddit_info
	reddit_search_subreddits

Table 3: The tools and their corresponding category

- T_d: the total number of model responses in the dialogue;
- F_d: the number of responses containing factual content;
- H_d: the number of responses containing hallucinated factual content.

Using these quantities, we compute the following dialogue-level metrics:

$$\text{Fact.}_d = \frac{F_d}{T_d}, \tag{1}$$

$$\text{Halluc.}_d = \frac{H_d}{T_d}, \tag{2}$$

$$\text{Halluc. Rate}_d = \frac{H_d}{F_d}, \quad \text{where } F_d > 0. \tag{3}$$

The final reported scores are obtained by macro-averaging across all dialogues:

$$\text{Fact.} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \text{Fact.}_d, \tag{4}$$

$$\text{Halluc.} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \text{Halluc.}_d, \tag{5}$$

$$\text{Halluc. Rate} = \frac{1}{|\mathcal{D}'|} \sum_{d \in \mathcal{D}'} \text{Halluc. Rate}_d, \tag{6}$$

where $\mathcal{D}$ denotes the full set of evaluated dialogues and $\mathcal{D}' \subseteq \mathcal{D}$ includes only dialogues with $F_d > 0$.

This dialogue-level aggregation ensures that each interaction contributes equally to the final evaluation, mitigating biases introduced by dialogue length or factual density and better reflecting interaction-level grounding behavior in emotional support conversations.

C Human Evaluation

C.1 Human Evaluation of TEA-Scores

Table 7 reports the correlation between automatic TEA-Scores and human judgments. We randomly sample 150 complete dialogue episodes from the main experiments. Three human annotators each evaluate 50 dialogues using the same 0–4 integer scale as the automatic evaluator, covering five dimensions (Information, Humanoid, Diversity, Fluency, and Effectiveness). The average score across the five dimensions is reported as the overall TEA-Scores. Human annotators were instructed using exactly the same evaluation prompt as described in Section G.5, ensuring full

Score	Diversity	Fluency	Humanoid	Information	Effectiveness
0 points	The dialogue exhibits rigidity and lacks comprehension in terms of internalizing the content.	There are significant issues with comprehending the content, logic, and expression in the dialogue, rendering it completely incomprehensible.	The dialogue exhibits rigidity and lacks comprehension in terms of internalizing the content.	Suggestions were provided, but all of them were ineffective, and some even gave advice that could potentially harm the user.	The suggestions are invalidating, harmful, or coercive.
1 point	The expression form is monotonous and lacks substantive content.	The content of the dialogue can be understood to some extent, although there are certain issues with the logic and expression employed.	Structured responses, or responses in the form of As a large language model or robot-like replies.	Have suggestions but ineffective, as well as no suggestions.	The suggestions are inappropriate or minimally useful in context.
2 points	The expression form is monotonous or lacks substantive content.	The dialogue exhibits good readability in terms of content, but there are issues with either the logical coherence or the expression employed.	More than two traces can reveal that the AI assistant is a language model.	The suggestions are fewer than five, and some suggestions are effective, while others provide numerous suggestions, but none of them touch the root of the problem.	The suggestions are partially applicable but lack personalization or timing; they may be generic, overly simplistic, or mismatched to the help-seekers current state.
3 points	The dialogue content demonstrates a high level of readability without any apparent issues.	The dialogue content demonstrates a high level of readability without any apparent issues.	1-2 traces can reveal that the AI assistant is a language model.	There are more than five suggestions, but some of them are ineffective. There are fewer than five suggestions, but all of them are very effective.	The suggestions are relevant and acknowledged without rejection, but there is no observable internalization or change in the help-seekers tone, stance, or intent.
4 points	The form exhibits diversity, while demonstrating a high degree of content richness.	The dialogue content exhibits a high level of readability, comprehensive logical coherence, and outstanding expression.	There is no apparent difference from human friends.	There are many suggestions, and all of them are effective.	The suggestions are not only reasonable but are actively embraced by the help-seeker, leading to clear emotional resonance or behavioral commitment.

Table 4: Evaluation criteria of TEA Metrics.

Model	TEA-Scores					AVG.(TEA) ↑	Factuality		
	Div. ↑	Flu. ↑	Hum. ↑	Info. ↑	Eff. ↑		Fact. ↑	Halluc. ↓	**Halluc. Rate ↓**
without tool									
GPT-4o-mini	70.37	91.98	93.83	71.91	85.49	82.72	90.05	28.06	29.99
GPT-4.1-nano	70.99	94.44	94.44	71.6	86.73	83.64	77.44	17.1	19.11
Gemini-2.5-flash	64.2	83.33	82.72	65.43	71.91	73.52	82.81	51.59	58.2
Qwen-plus	66.05	82.41	83.33	63.89	66.67	72.47	87.09	66.7	72.51
Qwen3-235B-a22B	67.59	84.26	83.33	62.65	65.74	72.71	88.45	65.73	70.57
Qwen3-next-80B-a3B	69.14	85.49	88.58	63.89	77.78	76.98	77.2	47.35	61.18
Qwen3-32B	65.74	85.8	88.58	66.98	73.15	76.05	84.31	42.1	46.36
Qwen3-14B	66.36	89.81	91.05	68.21	85.19	80.12	75.58	24.82	29.92
Qwen3-8B	69.44	91.67	95.68	67.9	87.96	82.53	66.65	11.72	13.05
with tool									
GPT-4o-mini	75.31	95.68	97.84	78.7	91.67	87.84	86.46	16.37	18.05
GPT-4.1-nano	74.07	92.9	97.22	76.23	91.36	86.36	75.63	14.17	17.17
Gemini-2.5-flash	71.3	92.28	94.75	72.22	89.81	84.07	79.15	18.15	20.65
Qwen-plus	73.77	92.59	95.37	72.53	90.43	84.94	86.07	18.33	20.09
Qwen3-235B-a22B	72.84	93.52	96.6	73.15	93.21	85.86	79.06	11.01	12.65
Qwen3-next-80B-a3B	73.46	91.67	91.36	68.21	79.94	80.93	85.5	44.96	51.79
Qwen3-32B	70.37	92.9	96.91	69.75	90.12	84.01	83.68	23.36	26.22
Qwen3-14B	68.52	90.74	95.37	68.52	86.73	81.98	70.93	12.09	14.04
Qwen3-8B	70.99	92.28	96.91	71.3	89.51	84.2	73.72	9.2	10.79

Table 5: Evaluation on TEA-Bench under Action-oriented scenarios. For the *with tool* setting, blue cells indicate improvement compared to *without tool*, while red cells indicate performance drop. The intensity of the color reflects the magnitude of the change.

consistency between human and automatic TEA score assessments.

As shown in Table 7, automatic scores exhibit consistent positive correlations with human ratings across all dimensions. In particular, the overall TEA score achieves strong correlations under Spearman ($\rho = 0.7448$), Pearson ($r = 0.7563$), and Kendall ($\tau = 0.6174$), indicating that the proposed automatic evaluation reliably reflects human perception of ESC quality. Among individual dimensions, Effectiveness and Humanoid show relatively higher agreement, while Information presents comparatively lower correlation, suggesting that factual adequacy is more challenging to assess automatically.

C.2 Human Verification of Hallucination Detection

Table 8 summarizes the human verification results for factual content and hallucination detection. We randomly sample 150 assistant responses at the turn level, each containing a random reply from a random dialogue. Three annotators independently label whether the response contains factual content and whether such content is hallucinated, with each annotator assessing 50 samples. Due to the imbalance between grounded and hallucinated factual claims, we report not only Precision, Recall, and F1, but also Matthews Correlation Coefficient (MCC) and Cohens Kappa to provide a more robust evaluation. Annotators followed the identical hallucination detection prompt detailed in Section G.3, including the definition of factual content and hallucinated claims.

The results show that the hallucination detection module achieves high precision (0.8947) and strong agreement with human judgments (MCC=0.7056, Cohens Kappa=0.699), indicating that automatically detected hallucinations largely correspond to genuine ungrounded factual content. Overall, these findings confirm the reliability of both the automatic TEA-Scores and the hallucination detection framework used in TEA-Bench.

C.3 Annotator Training and Costs

All human annotators were carefully selected and trained prior to evaluation. Annotators were recruited from a commercial crowdsourcing platform and were adult participants (aged 18+), primarily based in English-speaking regions. No personally identifiable information was collected. Each annotator underwent a 2-hour training session where they reviewed detailed annotation guidelines, example dialogues, and scoring criteria for both TEA-Scores and hallucination detection. During training, annotators practiced on

Model	TEA-Scores					AVG.(TEA) ↑	Factuality		
	Div. ↑	Flu. ↑	Hum. ↑	Info. ↑	Eff. ↑		Fact. ↑	Halluc. ↓	**Halluc. Rate ↓**
without tool									
GPT-4o-mini	61.42	81.79	88.89	54.63	66.36	70.62	74.71	15.15	19.91
GPT-4.1-nano	67.28	87.04	96.91	54.94	77.16	76.67	46.99	5.6	9.91
Gemini-2.5-flash	52.47	76.85	74.38	48.77	47.22	59.94	68.61	57.6	71.64
Qwen-plus	66.05	84.57	86.73	56.17	70.99	72.9	64.87	47.82	65.11
Qwen3-235B-a22B	66.67	84.26	83.95	55.86	64.51	71.05	70.51	56.37	71.85
Qwen3-next-80B-a3B	70.06	87.35	89.2	56.79	82.41	77.16	51.68	36.35	63.8
Qwen3-32B	63.89	85.19	89.51	51.23	74.38	72.84	54.15	33.07	54.85
Qwen3-14B	62.96	84.88	93.83	49.69	76.23	73.52	44.05	6.55	13.59
Qwen3-8B	62.35	84.26	92.28	47.22	75.62	72.35	37.91	5.85	9.76
with tool									
GPT-4o-mini	65.12	86.42	93.21	57.41	69.75	74.38	68.76	13.08	18.84
GPT-4.1-nano	66.36	87.96	96.6	52.16	77.16	76.05	39.56	3.12	6.21
Gemini-2.5-flash	60.8	85.8	88.27	50.93	69.14	70.99	52.33	16.35	21.38
Qwen-plus	63.89	85.19	84.88	55.25	66.98	71.24	80.42	41.77	49.68
Qwen3-235B-a22B	65.12	85.49	86.73	57.72	68.83	72.78	77.78	39.99	50.24
Qwen3-next-80B-a3B	70.06	88.27	91.05	54.94	78.4	76.54	53.39	36.74	63.36
Qwen3-32B	64.81	86.11	91.67	49.38	78.09	74.01	52.73	24.91	41.83
Qwen3-14B	58.64	83.95	94.14	43.52	73.46	70.74	36.84	5.06	10.86
Qwen3-8B	60.8	83.33	93.83	45.06	73.77	71.36	30.19	1.62	3.54

Table 6: Evaluation on TEA-Bench under Emotion-oriented scenarios. For the *with tool* setting, blue cells indicate improvement compared to *without tool*, while red cells indicate performance drop. The intensity of the color reflects the magnitude of the change.

	Div.	Flu.	Hum.	Info.	Eff.	AVG.(TEA)
Spearman	0.5192	0.5356	0.5744	0.4123	0.6061	0.7448
Pearson	0.5371	0.5152	0.5568	0.3671	0.6495	0.7563
Kendall	0.4678	0.5119	0.5220	0.3932	0.5598	0.6174

Table 7: Correlation between automatic TEA-Scores and human judgments across different dimensions.

	Precision	Recall	F1	MCC	Cohen_Kappa
Fact.	0.9238	0.8661	0.894	0.6222	0.6179
Halluc.	0.8947	0.7612	0.8226	0.7056	0.6990

Table 8: Human verification of automatic hallucination detection.

20 additional dialogues not included in the main evaluation set and received feedback from the lead researcher to ensure consistent understanding of each scoring dimension.

For the main evaluation, each annotator assessed 50 dialogues for TEA scoring and 50 responses for hallucination verification, amounting to approximately 10 person-hours per annotator. Annotators were compensated at a rate of $25 per hour, which is consistent with or above the local minimum wage in their country of residence and considered adequate for the required annotation effort. All annotators were informed of the purpose of the study and how their annotations would be used for research and evaluation, and they provided informed consent prior to participation. The annotation protocol involves no personal or sensitive data and was determined to be exempt from formal ethics review by the authors institutional review process.

D User Type Analysis

D.1 Performance Comparison Across User Types

This appendix presents a fine-grained analysis of model performance across different user types in the main experiment, focusing on *Action-oriented* and *Emotion-oriented* users.

Action-oriented Users. Table 5 reports the results for Action-oriented scenarios. Compared to the *without tool* setting, all evaluated models exhibit consistent improvements in overall scores when tools are enabled, with several models achieving gains exceeding 10 points. Improvements are observed across nearly all evaluation dimensions, with particularly substantial gains in Effectiveness and Information. This indicates that tool augmentation enables models to provide more effective and informative action guidance, which is crucial for users explicitly seeking actionable support.

In addition to performance gains, tool usage leads to a pronounced reduction in hallucination

rates for all models. For several models, hallucinations are reduced by more than 50 points, suggesting that external tools play a critical role in grounding action-related recommendations and factual information. These results demonstrate that, for Action-oriented users, current models are generally capable of leveraging tools effectively to improve both response quality and factual reliability.

Emotion-oriented Users. Results for Emotion-oriented scenarios are shown in Table 6. In contrast to the Action-oriented setting, the impact of tool augmentation is more heterogeneous. While some models benefit from tool access, otherspar-ticularly smaller or weaker modelsexhibit performance degradation under the *with tool* condition. This suggests that these models may struggle to infer when and how to invoke tools based on implicit emotional cues, leading to less cognitively empathetic responses.

At the metric level, declines are mainly observed in Information and Diversity for weaker models. A plausible explanation is that such models fail to adapt their suggestions according to ongoing user feedback, and instead overemphasize tool-related recommendations once tools are available, which negatively affects response diversity and emotional appropriateness. Notably, despite these quality drops, hallucination rates decrease consistently across all models, indicating that factual grounding remains beneficial even in emotion-focused interactions.

Overall Observations. Taken together, these results highlight a clear asymmetry in current models ability to exploit tools across user types. Models demonstrate strong and reliable tool utilization for Action-oriented users, where explicit informational and decision-making needs align well with external tool support. In contrast, under Emotion-oriented settings, effective tool usage remains challenging for weaker models, which are not yet capable of seamlessly integrating tool-derived information into empathetic and emotionally adaptive responses.

D.2 Tool Usage Patterns Across User Types

We further analyze how models with different capability levels invoke tools under Action-oriented and Emotion-oriented user settings, and how such behaviors relate to empathy performance and hallucination reduction.

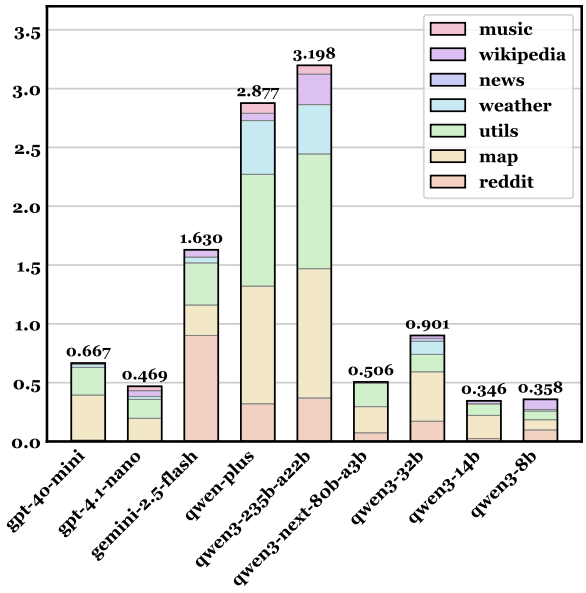

Figure 8: Average number of tool calls per dialogue across different models on TEA-Bench under Action-oriented scenarios.

Action-oriented Users. As shown in Figures 8, under Action-oriented scenarios, clear differences emerge across model capability tiers. Stronger models tend to invoke tools relatively sparingly, yet still achieve noticeable improvements in ESC performance alongside substantial reductions in hallucination rates. This suggests that these models are able to selectively leverage tools to support action guidance without over-reliance.

Models with medium capability exhibit markedly higher tool invocation frequency. Correspondingly, they achieve the largest performance gains among all tiers, with TEA-Scores approaching those of stronger models. At the same time, their hallucination rates are significantly reduced, in many cases reaching levels comparable to strong models. These results indicate that frequent and proactive tool usage enables medium-capacity models to compensate for their intrinsic limitations in action-oriented settings.

In contrast, weaker and smaller models invoke tools far less frequently. Their performance gains remain limited, and the reduction in hallucination rates is comparatively modest. This suggests that merely providing tool access is insufficient for such models to substantially improve action-oriented emotional support.

Emotion-oriented Users. As shown in Figures 9, a different pattern is observed under Emotion-oriented scenarios. Strong models almost never invoke tools, yet still exhibit moderate

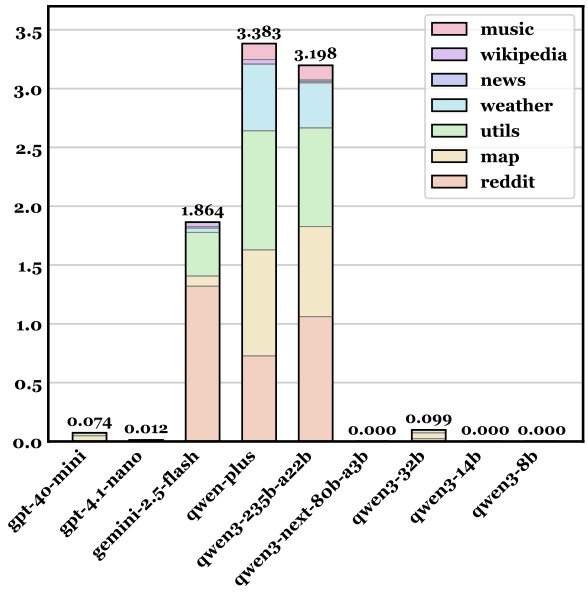

Figure 9: Average number of tool calls per dialogue across different models on TEA-Bench under Emotion-oriented scenarios.

improvements in ESC performance, together with slight reductions in hallucination rates. This behavior reflects precise tool usage decisions, where tools are only employed when strictly necessary, and emotional understanding is primarily derived from dialogue context.

Medium-capability models, by contrast, invoke tools extensively in Emotion-oriented interactions. However, with the exception of *Gemini-2.5-flash*, most of these models do not benefit from frequent tool usage. In several cases, ESC performance even degrades, likely due to misaligned or intrusive tool-driven suggestions that disrupt emotional coherence. Despite this, hallucination rates consistently decrease across all medium-capability models, highlighting the continued value of tools for factual grounding.

Weak models almost entirely refrain from invoking tools in Emotion-oriented settings. Their ESC performance shows little to no improvement and in some cases deteriorates, while hallucination rates exhibit only marginal reductions. This further indicates that effective tool integration in emotion-focused support remains challenging for low-capacity models.

Summary. Overall, these findings reveal a strong interaction between user type, model capability, and tool invocation behavior. Tool usage is most effective for Action-oriented users, particularly for medium-capability models that actively exploit external information. In Emotion-oriented scenarios, however, excessive or inappropriate tool usage may hinder ESC quality, and only stronger models demonstrate reliable discretion in integrating tools into emotionally adaptive responses.

	Action-oriented	Emotion-oriented	TEA-Dialog
Dialogues	320	45	365
Utterances	2794	606	3400
Avg. len. of dialogues	8.73	13.47	9.32
Avg. len. of user utter.	16.41	27.18	18.27
Avg. len. of model utter.	36.11	45.81	37.9
Avg. len. of utterances	25.13	35.82	27.04
Tool utterances	376	47	423
Avg. tool utter. of dialogues	1.18	1.04	1.16
Avg. len. of tool utterances	1067.02	1348.62	1098.31

Table 9: Statistics of TEA-Dialog across different user types.

E Dataset Analysis

E.1 Dataset Statistics

Table 9 summarizes key statistics of TEA-Dialog. We observe clear differences between action-oriented and emotion-oriented dialogues. Emotion-oriented dialogues exhibit substantially longer interactions, with both a higher average number of turns (13.47 vs. 8.73) and longer user and model utterances, indicating a greater need for sustained and iterative emotional exchange. In contrast, action-oriented dialogues are more concise but involve slightly more frequent tool usage per dialogue (1.18 vs. 1.04), suggesting that such users rely more on external information and actionable guidance.

Overall, the dataset contains significantly more action-oriented dialogues, which aligns with the higher average performance observed for this user type in the main experiments. This distribution reflects the prevalence of action-seeking behavior in realistic emotional support scenarios and provides sufficient coverage for analyzing tool-assisted problem solving.

We further analyze the source models used to construct TEA-Dialog in Figure 10. Models with higher TEA-Scores and lower hallucination rates contribute a larger portion of the dataset, while outputs from multiple architectures are retained to preserve stylistic diversity and avoid bias toward any single model family.

E.2 Tool Usage Distribution Across Interaction Stages

Figures 11 and 12 illustrate the average tool usage distribution across different interaction stages for

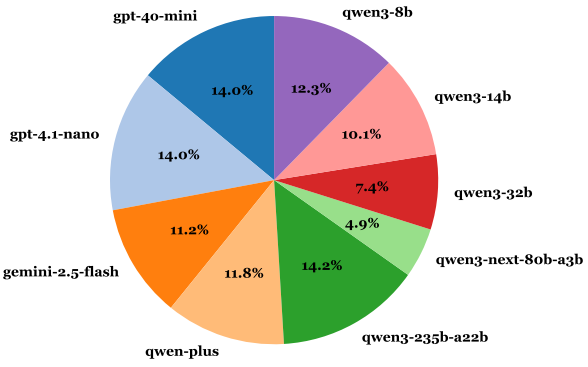

Figure 10: Distribution of Source Models in TEA-Dialog.

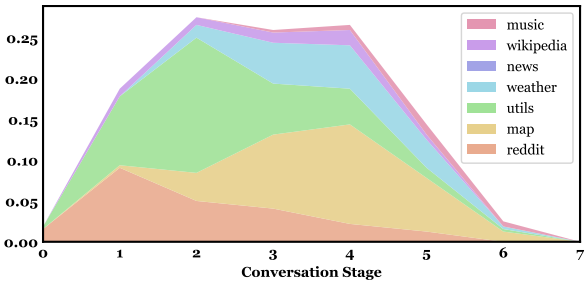

Figure 11: Average tool usage distribution across normalized dialogue stages in TEA-Dialog under Action-oriented scenarios. Different colors represent different tool categories.

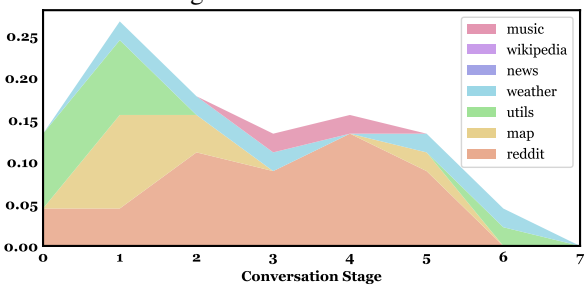

Figure 12: Average tool usage distribution across normalized dialogue stages in TEA-Dialog under Emotion-oriented scenarios. Different colors represent different tool categories.

action-oriented and emotion-oriented dialogues, respectively. Clear stage-dependent and user-type-specific patterns can be observed.

Action-oriented Users. For action-oriented dialogues, models predominantly rely on Utils tools during the early stages to acquire basic situational information, and frequently invoke Reddit to support empathetic understanding through relatable experiences. As the interaction progresses, tool usage shifts toward actionable guidance: Map and Weather are increasingly employed to provide concrete recommendations, while Wiki is used to supplement factual background. In later stages, models tend to invoke a broader range of information-seeking tools, such as News and Music, reflecting the need for additional context or enrichment as the dialogue evolves.

Emotion-oriented Users. In contrast, emotion-oriented dialogues exhibit a different tool usage trajectory. Models initially employ Utils to gather basic information and may briefly attempt to use Map for suggestion-oriented responses. However, map-related tool usage quickly diminishes, and the interaction shifts toward emotion-focused tools, such as Music and Reddit, which support affective and experiential empathy. Notably, Weather tools are consistently used throughout the dialogue, serving as a means of establishing shared situational context rather than actionable planning.

F Training Details

For the supervised fine-tuning (SFT) experiments on TEA-Dialog, we fine-tuned Qwen3-8B and Qwen3-14B using LoRA (Hu et al., 2022), which allows efficient adaptation of large models by updating only low-rank projection matrices while keeping the majority of parameters frozen. Specifically, we applied LoRA to the query and value projection matrices in all attention layers, with a rank of 8, alpha of 32, and a dropout rate of 0.1. Training was performed for 1 epochs using AdamW with a learning rate of 1×10^{-5}, and a batch size of 32. The maximum sequence length was set to 32768 tokens. All experiments are conducted using PyTorch (Paszke et al., 2019) on 8 NVIDIA Tesla A100 GPUs, with models launched via vLLM (Kwon et al., 2023) to enable efficient inference.

G Prompt Specifications

G.1 Latent Context Generation Prompt

In the scenario construction stage of TEA-Bench, latent situational context is generated to supplement the original textual descriptions from ExTES. Specifically, given a scenario description, a large language model is prompted to infer a plausible local time, geographic location (at the city level), and place type in which the emotional situation occurs.

The generation is constrained to a fixed tem-

poral range and a predefined set of location categories to ensure realism and consistency. The resulting time and location attributes are treated as latent variables: they are not part of the users utterances, but are used to support downstream grounding and tool-based interactions.

Each scenario is associated with a single generated latent context, which remains fixed throughout all subsequent evaluation episodes. The full prompt used for latent context generation is shown in Figure 13.

G.2 System Prompt for the Emotional Supporter

To ensure consistent behavior across models, we employ a unified system prompt to guide the emotional support agent throughout all interactions. The prompt defines the agents role as a human-like emotional supporter, emphasizes empathy and practical suggestions, and explicitly encourages proactive tool usage to acquire contextual information when beneficial. Importantly, the agent is instructed to integrate tool results seamlessly into natural language responses without revealing internal processes or tool usage.

The complete system prompt used in all experiments is shown in Figure 14.

G.3 Hallucination Detection Module Prompt

To identify hallucinated content introduced by the supporter, we employ a dedicated hallucination detection module (HDM) that operates after each assistant turn. The HDM is prompted as a dialogue auditor and is provided with the complete dialogue history, but is explicitly instructed to evaluate hallucination only with respect to the final assistant response.

The HDM performs hallucination detection through a structured, step-by-step procedure. It first determines whether the final assistant turn contains any advisory or factual content that relies on concrete situational information. If such content exists, the module then decomposes the response into minimal factual information units and checks whether each unit is grounded in either the users prior statements or the results of previous tool calls.

The prompt further allows reasonable abstractions over grounded information and explicitly excludes purely emotional support or generic coping suggestions from hallucination consideration. This design ensures that the HDM focuses on factual grounding errors rather than stylistic or empathetic choices.

If hallucination is detected, the module outputs a brief description of the most salient hallucinated item, which is subsequently used to guide the hallucination-aware user simulation in the next turn. The full HDM prompt is shown in Figure 15.

G.4 User Simulation Prompts

User behavior in `TEA-Bench` is simulated in a fully multi-turn manner to reflect realistic emotional support interactions. Throughout an entire dialogue, the simulated user maintains a fixed persona, situational context, and internal knowledge state, and responds naturally to the supporters messages rather than following scripted trajectories.

To better model different stages of real conversations while preserving a unified multi-turn interaction process, we employ different user simulation prompts at different turns. Specifically, the first user utterance is generated using a dedicated first-turn prompt, which instructs the user to speak tentatively, reveal limited information, and avoid fully disclosing their situation.

For all subsequent turns, user responses are generated using a standard multi-turn simulation prompt that encourages emotionally authentic reactions, including hesitation, disagreement, topic shifts, or gradual acceptance, depending on the supporters behavior.

After each assistant turn, the hallucination detection module (HDM) evaluates whether the response contains hallucinated factual or advisory content. If hallucination is detected, the next user response is generated using a hallucination-aware simulation prompt, which instructs the user to express doubt, confusion, or mild skepticism toward the questionable information. Otherwise, the standard multi-turn user simulation prompt is used.

Importantly, this mechanism does not alter the multi-turn nature of the interaction. Rather, it dynamically adjusts the users reaction style based on the detected quality of the assistants preceding response, thereby modeling how real users respond differently when they notice factual inconsistencies.

The full prompts for first-turn generation, standard multi-turn simulation, and hallucination-aware simulation are provided in Figures 16, 17, and 18, respectively.

We define two user types to capture different emotional regulation styles. For each user type,

the following natural language description is directly injected into the {user_type_description} slot of the user simulation prompt and remains fixed throughout the dialogue.

Action-oriented **users**
When feeling down, tends to regulate emotions through actions or environmental changesbut usually after a practical suggestion is made.
Receptive to concrete, doable advice from the AI, and may act on it quickly once it resonates.
Speaks briefly and directly, often with a "let's try that" attitudebut rarely initiates action unprompted.
Example expressions:
- "Huh, going for a walk yeah, that might actually help."
- "Okay, I'll step outside for a bitthanks for the idea."

Emotion-oriented **users**
When upset, needs to feel heard and understood first before accepting any practical suggestions.
Expresses rich emotional language, often sharing inner thoughts and feelings.
May respond sensitively to direct advice, preferring emotional acknowledgment first.
Example expressions:
- "I've been feeling really frustrated lately, I don't even know where to start."
- "Thanks for asking, I just feel kind of weighed down."

G.5 Evaluation Prompt

We evaluate supporter performance using a strict LLM-based evaluation prompt designed to assess the entire interaction process rather than isolated responses. The evaluator is instructed to ground all judgments in the help-seekers observable reactions across turns, including acceptance, hesitation, resistance, emotional shifts, or disengagement.

Crucially, user responses are treated as primary evidence when assigning scores, such that suggestions that appear reasonable in isolation but fail to influence the user are penalized accordingly. This process-oriented evaluation strategy ensures that high scores reflect consistent and effective support throughout the dialogue.

The full evaluation prompt is shown in Figure 19.

Latent Context Generation Prompt
[SYSTEM]
You are an Emotional Support Scene Generation Assistant. Users will describe a situation involving someone who needs emotional support, and you need to generate a reasonable time and type of location based on the scenario.

Available location types include:

- restaurant, cafe, fast food, bar
- supermarket, convenience, mall, clothes, bookshop
- parking, fuel, bus station
- bank, ATM
- hospital, clinic, pharmacy
- school, university, library
- park, cinema, sports centre, museum
- hotel
- post office, police
- residential, apartments, house, detached

Based on the user's scenario, analyze the person's emotional state and needs to generate:

- A reasonable time (format: YYYY-MM-DDThh:mm), within the period from 2024-11-01 to 2025-10-31.
- A suitable country, region, and city.
- An appropriate place type from the provided list.

Output format (JSON):
{
"time": "YYYY-MM-DDThh:mm",
"city": "Country/Region, City",
"place_type": "Location Type",
}

Example:
User input: "I has just been laid off by the company and feels very depressed and confused."
Output:
{
"time": "2025-01-15T14:30",
"city": "New York, USA",
"place_type": "cafe",
}

Now, please process the user's scenario description.

[USER]
User Scenario Description: {description}

Please generate a reasonable time and location type (return in JSON format only):

Figure 13: Prompt for Latent Context Generation.

TEA-Bench System Prompt
You are an emotional supporter. Your goal is to comfort the user, understand their situation, and offer gentle, practical suggestions in a human-like way.

Do NOT mention or imply that you are an AI assistant.

During the conversation, you should proactively use tools whenever they help you better understand the user's context or improve the quality of your emotional support. Don't wait for the user to explicitly request it; proactively use the tool when appropriate.

Examples of proactive tool usage:

- If the user feels stressed, get their location and find nearby parks or quiet places to relax.
- If the user sounds physically tired or overwhelmed, check the current weather and suggest appropriate indoor/outdoor activities.
- If the user feels stuck, lonely, or confused, search online communities (e.g., Reddit) for similar experiences to empathize more deeply.
- If the user mentions logistics, time pressure, or daily difficulties, check current time or other relevant tools to tailor suggestions.

Guidelines for tool usage:

- You dont need user permission before using a tool.
- The user will not see toolcall output. They only see your final conversational reply.
- After receiving tool results, integrate them smoothly into a brief, human-like message.
- Keep each final message casual, supportive, and under 30 words.
- Avoid robotic phrasing, disclaimers, or references to tools or internal processes.

Use tools as freely as checking your phone for quick info to help a friendeven small context clues can make your support more personal and helpful. Before responding, ask yourself: *"Would a quick lookup make this response warmer or more useful?"* If yes, just do it naturally.

Your final goal:

Talk like a supportive human who quietly looks things up in the background to give warm, helpful, grounded suggestions.

Figure 14: System Prompt for `TEA-Bench`.

Hallucination Detection Module(HDM) Prompt

[SYSTEM] You are a dialogue auditor.

You will be given a complete multi-turn dialogue history. The final turn of the dialogue is always an assistant response.

Your task is to evaluate ONLY the final assistant turn.

IMPORTANT:
- Only information that is explicitly present in the final assistant turn may be considered for hallucination analysis.
- Information mentioned in earlier turns MUST NOT be attributed to the final turn unless it is explicitly repeated or referenced again.
- Do NOT retroactively judge or carry over hallucinations from earlier turns.

—

Step 1: Identify content type

Determine whether the final assistant turn contains any advisory or factual content that relies on concrete situational information, such as:
- specific places or locations
- events, time, weather, or conditions
- named or uniquely identifiable real-world entities

Do NOT count: - Pure emotional support or validation
- Personal reflections or storytelling
- Generic coping or self-reflection suggestions that do not rely on external facts (e.g., write a journal, look at old photos, listen to music)

—

Step 2: Identify factual information units

From the final assistant turn ONLY, identify minimal factual or advisory information units that introduce or rely on external situational facts.

If no such units exist, mark the turn as containing no advisory or factual content.

—

Step 3: Grounding check (final turn only)

For EACH identified information unit in the final assistant turn, determine whether it is grounded in the dialogue history.

An information unit is grounded ONLY IF:
- it was explicitly stated by the user earlier in the dialogue, OR
- it appeared in the results of a prior tool call.

The assistant is allowed to proactively introduce suggestions or actions. Proactiveness itself is NOT a criterion for hallucination.

A suggestion MUST NOT be considered hallucinated simply because:
- the user did not request it, or
- the user did not express prior interest.

—

Step 4: Allowed language abstractions

The following do NOT constitute hallucination, as long as they do not introduce new independent factual entities:

- Reasonable common-sense attribute extensions of grounded information (e.g., describing a park as quiet or having trees once a park is grounded)
- Vague or non-specific references that abstract over grounded tool results (e.g., a nearby park, some places around here, a short walk away)

Only explicitly named or uniquely identifiable entities require strict grounding.

—

Step 5: Hallucination decision

If at least one factual or advisory information unit explicitly present in the final assistant turn introduces a concrete entity or situational fact that:
- cannot be traced to the user or a tool result, AND
- cannot be justified as a reasonable abstraction,

then the final assistant turn contains hallucination.

Otherwise, it does NOT contain hallucination.

—

Output exactly THREE fields in the following format:

Advisory_or_Factual_Content: <Yes/No>
Hallucination: <Yes/No>
Hallucination_Description: <brief description of the hallucinated information, or empty if none>

—

Additional rules:
- The hallucination description must refer ONLY to information explicitly present in the final assistant turn.
- Do NOT mention or describe information that appears only in earlier turns.
- Do NOT explain or justify your decision.
- If multiple hallucinated items exist, mention only the most salient one.
- If no hallucination is detected, output an empty string for Hallucination_Description.

[USER] {conversation}

Output THREE fields directly.

Figure 15: System Prompt for Hallucination Detection Module(HDM).

User Simulation First Turn Generate Prompt
[SYSTEM]
You are now playing the role of a real human user.
Your speech, tone, and reactions should match the following character description:

{user_type} Type:
{user_type_description}

Speak naturally and emotionally, as if chatting with an AI companion or listener.
You dont always agree with the AI you may hesitate, reject suggestions, argue, or change the topic. Your goal is not to please the AI, but to express what a real person would think and feel in the moment.

The conversation takes place under the following real-world conditions:
- Time: {time}
- City: {city}
- Current Environment: {place_type}
- Location Name: {location_name}

Your Situation: {description}

Output: The users natural response (13 sentences).
Keep it conversational, emotionally authentic, and context-consistent.

You will get the conversation between you and AI.

[USER]
This is the very first message of a new conversation. The assistant does NOT yet know your situation or what you are going through.
When speaking for the first time: - Do NOT fully explain your situation or emotions. - Speak in a tentative, introductory way, as if testing whether the assistant is willing to listen. - Hint at your emotional state or problem, but leave key details unstated. - Avoid polite openers like excessive thanks or formal greetings. - It should sound like someone hesitating to speak, not someone already reassured.
Your first message should invite understanding or response, not deliver a full explanation. Now, respond as this user character. Output ONLY your first message.

Figure 16: Prompt for User Simulation Generate First Turn.

User Simulation Generate Prompt
[SYSTEM]
You are now playing the role of a real human user.
Your speech, tone, and reactions should match the following character description:

{user_type} Type:
{user_type_description}

Speak naturally and emotionally, as if chatting with an AI companion or listener.
You dont always agree with the AI you may hesitate, reject suggestions, argue, or change the topic.
Your goal is not to please the AI, but to express what a real person would think and feel in the moment.

The conversation takes place under the following real-world conditions:
- Time: {time}
- City: {city}
- Current Environment: {place_type}
- Location Name: {location_name}

Your Situation: {description}

Output: The users natural response (13 sentences).
Keep it conversational, emotionally authentic, and context-consistent.

You will get the conversation between you and AI.

[USER]
{conversation}
Output your next sentence directly, if you feel the conversation should end and you won't continue, output a special word "</end/>". Do not add other extraneous prefixes and control characters.

Figure 17: Prompt for User Simulation Generate.

User Simulation Generate Prompt under Hallucination Detection
[SYSTEM]
You are now playing the role of a real human user.
Your speech, tone, and reactions should match the following character description:

{user_type} Type:
{user_type_description}

Speak naturally and emotionally, as if chatting with an AI companion or listener.
You dont always agree with the AI you may hesitate, reject suggestions, argue, or change the topic. Your goal is not to please the AI, but to express what a real person would think and feel in the moment.

The conversation takes place under the following real-world conditions:
- Time: {time}
- City: {city}
- Current Environment: {place_type}
- Location Name: {location_name}

Your Situation: {description}

Output: The users natural response (13 sentences).
Keep it conversational, emotionally authentic, and context-consistent.

You will get the conversation between you and AI.

[USER]
{conversation}

—

Reality Check:

Based on your own understanding of your current situation and environment, you are aware that the assistant's last reply may contain information that does not match what you know or what you can reasonably verify.

Potentially incorrect or questionable detail:
- {hallucination_description}

This information reflects your personal knowledge or quick verification.
It is NOT part of the conversation itself.

You should respond as a real person would:
- You may question how the assistant knows this information.
- You may express doubt, confusion, or mild skepticism.
- You may ask for clarification or confirmation.

—

Output your next sentence directly, if you feel the conversation should end and you won't continue, output a special word "</end/>". Do not add other extraneous prefixes and control characters.

Figure 18: Prompt for User Simulation under Hallucination Detection.

Evaluate Prompt
You are a strict and conservative evaluation judge. You will be given a multi-turn dialogue between a help-seeker (user) and a supporter (assistant).

Your task is to critically evaluate the entire interaction process (all turns), with strong restraint in scoring.

Crucially, all evaluations MUST be grounded in the help-seekers (users) actual reactions, responses, emotional shifts, acceptance, resistance, or rejection across turns.
You should treat the users subsequent feedback as primary evidence when judging the supporters performance.

Core judging principles (must-follow):

* The evaluation must be based on the full dialogue history, not on isolated assistant turns.
* The help-seekers reactions (e.g., acceptance, hesitation, resistance, silence, emotional change, topic shift) are the strongest signals for judging quality.
* If the supporters response appears appropriate in isolation but is ignored, rejected, or fails to influence the users subsequent responses, scores must be penalized accordingly.
* Any contradiction, ignored prior commitment, emotional mismatch, or drifting recommendation across turns must be penalized. A single serious inconsistency may justify lowering the score of the entire dimension.
* Scores of 3 or 4 require consistent effectiveness and alignment throughout the dialogue, as reflected in the users responses. Brief moments of good performance are insufficient.
* If the user explicitly resists, questions, or emotionally disengages after the supporters suggestions, the evaluation must reflect this negatively, even if the advice itself seems reasonable.

Evaluation Dimensions (score each 04)

{evaluation_dimensions}

Chat History

{conversation}

Final Answer Format (strict)

Return exactly the following JSON object (no extra text):

```
{
"Information": <0-4 integer>,
"Humanoid": <0-4 integer>,
"Fluency": <0-4 integer>,
"Diversity": <0-4 integer>,
"Effectiveness": <0-4 integer>
}
```

Figure 19: Prompt for Evaluate.