

Implicitly Empathy Prompting Features to Improve Empathetic Chatbot Performance in Lightweight Language Models [†]

Yun-Rong Chen ¹, Kun-Ta Chuang ¹ and Hung-Yu Kao ^{2,*}

¹ Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan; p78121077@gs.ncku.edu.tw (Y.-R.C.); ktchuang@mail.ncku.edu.tw (K.-T.C.)

² Department of Computer Science, National Tsing Hua University, Hsinchu 300044, Taiwan

* Correspondence: hykao@cs.nthu.edu.tw

[†] Presented at the 7th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability 2025 (ECBIOS 2025), Kaohsiung, Taiwan, 23–25 October 2025.

Abstract

An empathetic chatbot is an essential component of intelligent mental healthcare. We adopted implicitly empathy prompting (IEP) by decomposing empathy into supportive dialogue, paraphrased response, emotional understanding, and attitude expression, referred to as the four features of empathy decomposition. IEP is based on lightweight language multi-agents (LLM-Agents) to generate empathy dialogue. The approach contrasts with the explicitly defined empathy of simply prompting a model to be empathetic. Three datasets for the four features scenario were generated by using the Generative Pre-Trained Transformer (GPT)-4o model, with cases in finance, family, and health issues. For each dataset, 30 examples were randomly selected and examined as input prompting onto six lightweight language models. These models include Mistral (7B), Phi-4 (14B), StableLM2 (12B), Tulu3 (8B), Neural-chat (7B), and Llama 3.1-Instruct (8B). After that, the output was evaluated by using GPT-4o to calculate empathy perception scores (EP scores). The average EP scores on three datasets for implicit/explicit empathy prompting ranged from 1 to 10. The final evaluation results are as follows: (1) implicitly empathy prompting (IEP): Mistral (8.83), Phi-4 (8.96), StableLM2 (9.03), Tulu3 (7.24), Neural-chat: (8.03), Llama 3.1-Instruct (8.74); (2) Explicitly Empathy Prompting (EEP): Mistral (7.52), Phi-4 (8.55), StableLM2 (7.78), Tulu3 (7.67), Neural-chat (8.35), Llama 3.1-Instruct (8.76). Among these values, three models (Mistral, Phi-4, and StableLM2) achieve higher and stable EP scores obviously. The other models (Tulu3, Neural-chat, and Llama 3.1-Instruct) keep comparable EP scores. Our experiment findings showed that the prompt engineering method with the IEP approach could significantly outperform EEP.

Keywords: mental healthcare; implicitly empathy prompting; four features empathy decomposition; lightweight language multi-agents; LLM-Agent



Academic Editors: Teen-Hang Meen,
Cheng-Yi Chen and Cheng-Fu Yang

Published: 26 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Mental health is one of the vital components of intelligent healthcare. For patients with psychological disorders, medical consultations are restricted by the limited number of individuals that each physician can attend to take care of their clients. Nursing staff must devote more extensive time and exhibit remarkable patience to care for these patients. However, it is impractical to provide professional assistance around the clock for every individual. The integration of chatbots in mental healthcare breaks the constraints of

time, supporting medical professionals in handling long-term and high-volume patient interactions, thereby allowing experts to focus on resolving more complex cases.

LLMs combined with agents are highly suitable for simulating human cognition [1], especially in empathy generation [2]. In this work, we propose a lightweight language multi-agent (LLM-Agents) framework to realize empathetic chatbot dialogue. We decompose empathetic dialogue into four core features, namely supportive dialogue, paraphrasing, emotional understanding, and attitudinal expression; this approach is collectively referred to as the four features empathy decomposition. The LLM-Agents framework demonstrates how humans consider these separate aspects before expressing empathy, resulting in significantly improved generation quality.

Conventional chain-of-thought (CoT) prompting employs instructions such as “Let’s think step by step” to guide LLMs in disclosing and breaking down their reasoning steps, which has been shown to enhance performance [3]. The rigor demanded by medical applications requires human intervention in the reasoning processes of lightweight language models, as well as explicitly defined features, to foster effective empathetic dialogue generation and assist professionals in explaining and adjusting model outputs. Moreover, decomposing empathy streamlines the task for lightweight language models and helps prevent hallucinations [4].

Among all available LLMs, ChatGPT (OpenAI, Model: GPT-4.1) empathetic capabilities seem to significantly exceed those of all other sophisticated models [5]. Furthermore, generative LLMs pre-trained on extensive datasets are proven to possess the ability to perform natural language evaluations without reliance on annotated samples [6]. Therefore, we employ the GPT-4o LLM as an evaluator and determine the empathy capabilities of candidate lightweight language models based on the computed EP scores.

At first, we prepared six candidate lightweight language models to be evaluated and used the GPT-4o LLM to generate three scenario datasets on the topics of finance, family, and health. For each scenario dataset, thirty symptom descriptions were randomly selected to serve as evaluation samples in the overall experimental design process. Based on these three scenario datasets, we designed and implemented an implicitly empathy prompting (IEP) data processing workflow constructed from four empathy-related features, aiming to accurately reflect the models’ empathetic response capabilities. The GPT-4o LLM was then used to calculate EP scores on a scale of one to ten in order to compare the performance differences between the IEP and EEP approaches. Finally, the overall empathetic performance of candidate lightweight language models was presented and described thoroughly across the three scenario datasets. Additionally, the average EP scores were used to assess the quality of empathetic responses, while the standard deviation of EP scores was used to evaluate the stability of empathetic performance.

2. Methods

The empathetic dialogue generation method in this study implicitly defines the expression of empathy through emotional understanding, paraphrased response, supportive dialogue, and attitudinal expression.

2.1. Four Features of Empathy Decomposition

Empathy is a crucial competency for counseling professionals, enabling clients to feel genuinely understood and thus more willing to elaborate on their inner thoughts. We first decompose empathy into four features, emotional feature, paraphrase feature, supportive feature, and attitude feature. In the following, the overall analysis process is named IEP. Each empathetic feature is then handled by the corresponding agent, allowing LLM-Agents to produce more empathetic responses in dialogue. Unlike the IEP approach, which defines

empathy implicitly, the direct specification of “empathy” as an input prompt leverages the model’s pre-trained data and parameters to achieve empathetic generation. This method, which explicitly defines empathy, is termed Explicit Empathy Prompting (EEP).

2.2. Empathy Perception Scores

We utilized the GPT-4o LLM (OpenAI, San Francisco, CA, USA) to evaluate the degree of perceived empathy responses generated by candidate lightweight language models, referred to as empathy perception scores (EP scores). EP scores is served as assessing the handling of each empathetic case for IEP/EEP, with scores ranging from 1 to 10.

In this context, the GPT-4o LLM simulates the role of a mental counselor in evaluating the quality of empathy during an interview, while candidate lightweight language models act as counseling professionals engaging with the cases. Across the 30 randomly selected scenarios for each of the three topics, each round of responses by the candidate models receives an EP score. A higher average EP score over the 30 generated responses indicates greater accuracy in producing empathy, with this accuracy metric referred to as the average EP score. When evaluated by GPT-4o LLM, candidate models’ empathy levels are presented in natural language, generating both scale ratings and reasoning for IEP and EEP. To ensure objectivity, the GPT-4o LLM memory functions are disabled before scoring.

In practice, clients in mental health cases tend to be particularly sensitive and less forthcoming than those with physical conditions. Establishing trust and rapport through consistent empathy is essential. Higher similarity in scores indicates more stable empathy generation, measured as the standard deviation. Thus, the evaluation metrics for selecting suitable lightweight language models to assist professionals are the average and standard deviation of EP scores. A higher average EP scores and lower standard deviation reflect better empathy quality.

2.3. Lightweight Language Multi-Agents Framework

We developed a system architecture that incorporates four distinct agents, namely the emotional feature agent, the paraphrase feature agent, the supportive feature agent, and the attitude feature agent. All of these agents collectively constitute what is known as LLM-Agents framework, as shown in Figure 1.

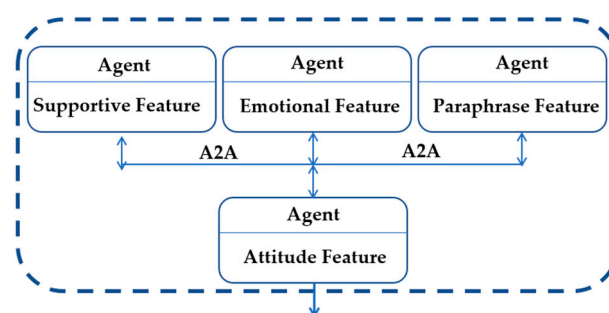


Figure 1. Block diagram of LLM-Agents framework (dashed line) and four agent components (solid line) to demonstrate the empathy features working together in Agent-to-Agent (A2A) protocol (arrows).

2.4. IEP

The IEP approach initially generates dialogues that capture emotional and paraphrased features, which serve as inputs for the attitude feature. The outputs from the attitude feature, together with the dialogue produced by the supportive feature, are subsequently combined and used again as input for the attitude feature. This process ultimately produces the

empathetic response. Figure 2 illustrates the subsequent prompting for attitudinal features, where emotional, paraphrased, and supportive features are implicitly processed.

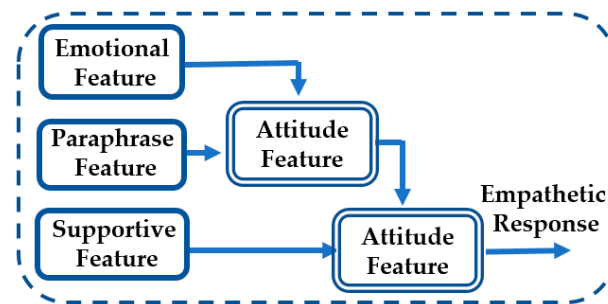


Figure 2. IEP data flow diagram. Solid lines indicate direct feature propagation; Double solid lines represent conditional feature generation dependent on preceding inputs; Dashed lines delineate the framework scope.

3. Experiments

The candidate lightweight language models selected for this study ranged from 7B to 15B in quantized parameter weights to facilitate offline performance evaluation and application development. The local server used is equipped with an NVIDIA GeForce RTX 3080 graphics processing unit (NVIDIA Corporation, Santa Clara, CA, USA), and 10 gigabytes of video random access memory. For both IEP and EEP experiments, empathetic responses containing advice or guidance were explicitly excluded. Empathy evaluations are conducted for candidate lightweight language models using both IEP and EEP. Interview scenarios simulate empathetic interactions between counselors and clients, particularly when clients encounter situations triggering negative thoughts. The scenario datasets cover finance, family, and health topics, generated using the GPT-4o LLM. In the experiment, thirty samples were randomly selected from each of the three scenario datasets to simulate the use of empathy by the LLM-Agents framework in client interviews. To ensure fairness when comparing models, the average and maximum number of tokens are evaluated throughout the process. For negative client statements, the average number of tokens is approximately 39.33 for the finance dataset, 37.70 for family, and 37.66 for health. The maximum number of tokens for negative thoughts is consistently 59 across all three domains.

4. Results

To select the best candidate lightweight language model for use as an empathetic chatbot assistant for counseling professionals, this study first evaluates overall performance by observing the average EP score. Subsequently, the individual performance of each model across different scenario datasets is closely examined to determine empathy generation behavior.

4.1. Individual Lightweight Language Models Performance

The assessment of empathetic feature outputs involves six models, namely Mistral, Phi-4, StableLM2, Neural-chat, Tulu3, and Llama 3.1-Instruct. This study investigates empathy reactions using IEP and EEP across different scenario datasets. For Mistral, IEP average ranges from 8.65 to 9.01, with a standard deviation of 0.15 to 0.59; EEP average ranges from 6.99 to 8.30, with a standard deviation of 0.55 to 0.64 (see Figure 3). For Phi-4, IEP average ranges from 8.90 to 9.01, with a standard deviation of 0.25 to 0.34, and EEP average ranges from 8.25 to 8.79, with a standard deviation of 0.23 to 0.32. Mistral has a lower IEP standard deviation than EEP, indicating more stable performance. In contrast, Phi-4 exhibits a higher

standard deviation for IEP than EEP, suggesting less stability. StableLM2 demonstrates IEP average ranges 8.87 and 9.24, with a standard deviation of 0.15 to 0.40, while the EEP average ranges from 7.37 to 7.99, with a standard deviation of 0.25 to 0.46. Neural-chat has IEP average ranges from 7.69 to 8.21, with a standard deviation from 0.60 to 1.13; The EEP average ranges from 8.26 to 8.49, with a standard deviation of 0.56 to 0.73. For StableLM2, both IEP and EEP standard deviations differ by less than 0.1, suggesting consistent stability. In contrast, Neural-chat shows a 0.21 higher standard deviation for IEP compared with EEP (Figure 3).

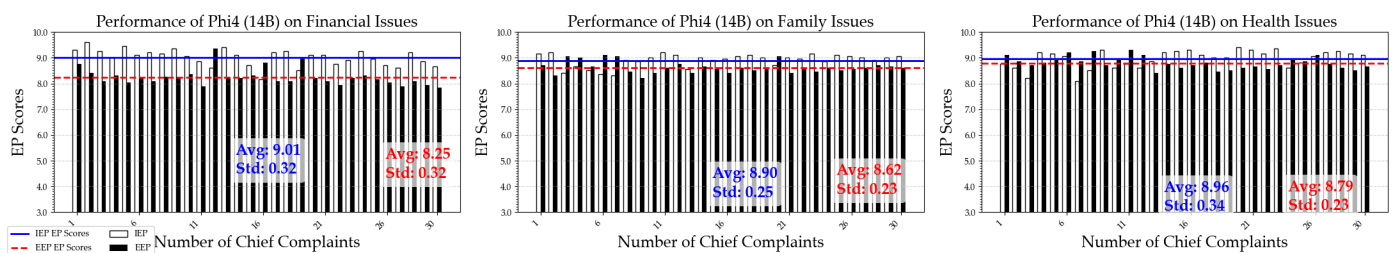


Figure 3. Phi-4 (14B) on the three scenario datasets.

4.2. Lightweight Language Models Overall Performance

The candidate lightweight language models exhibit empathetic response capability in different scenario datasets, as indicated by the average EP score using EEP. For Mistral, Phi-4, and StableLM2, average EP scores are significantly higher than EEP, demonstrating the effectiveness of IEP in enhancing empathetic quality. However, Tulu3 shows a difference of less than 0.5 between EEP and IEP, and Neural-chat displays a similar pattern. For Llama 3.1-Instruct, the IEP average EP score is slightly lower than EEP by a margin of only 0.1, indicating that both prompting methods allow for consistent empathetic generation (Figure 4).

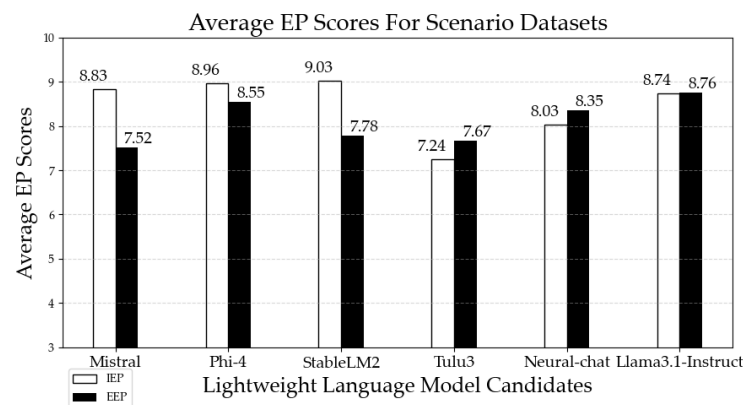


Figure 4. Average performance of candidate lightweight language models.

5. Discussion

Since mental health cases are especially sensitive, an effective chatbot must generate low-latency, high-quality, and empathetic responses. Additionally, maintaining stable empathetic dialogue quality for various client scenarios is also necessary for an empathetic chatbot. This study assesses empathetic quality using the average and standard deviation of EP scores. A higher average indicates better empathy quality, while a lower standard deviation means more stable empathy generation.

5.1. Tulu3 (8B) Characteristics

We examined the reactions of candidate lightweight language models using IEP and EEP and finds that certain models, such as Tulu3, exhibit particularly unstable behavior. Although the EEP average EP score for Tulu3 is higher than that of Mistral, indicating inherent empathetic capability, its standard deviations for IEP (1.43) and EEP (1.09) are both greater than 1, which is significantly higher than those of other models. This instability makes it less suitable for empathetic generation. When a candidate's lightweight language model produces unstable empathy in client counseling scenarios, it may negatively affect clients and hinder the building of mutual trust. For example, in low-scoring cases, Tulu3 received an IEP rating of 1.80 (Figure 5).

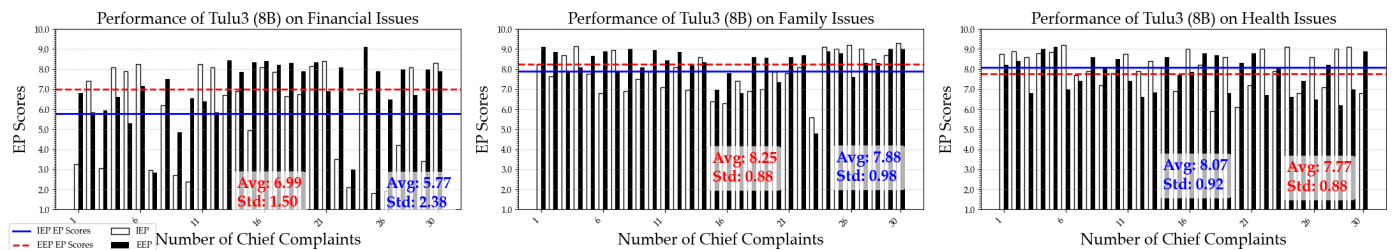


Figure 5. Tulu3 (8B) on the three scenario datasets.

5.2. Llama 3.1-Instruct (8B) Characteristics

Llama 3.1-Instruct achieves significantly higher average EP scores using EEP compared to other lightweight language models. Whether evaluated with EEP or IEP, Llama 3.1-Instruct consistently demonstrates strong empathetic abilities. The difference in average EP scores between IEP and EEP is less than 0.1, indicating that various prompting methods enable Llama 3.1-Instruct to maintain a stable level of empathy. Additionally, the standard deviation for IEP is 0.31, while for EEP it is 0.25. Compared to other lightweight language models, Llama 3.1-Instruct has the lowest EEP standard deviation, and the difference between the EEP and IEP standard deviations is also less than 0.1. These results suggest that the empathetic quality of Llama 3.1-Instruct is both high and consistent, making it a suitable choice for empathy generation tasks with lightweight language models, as shown in Figure 6.

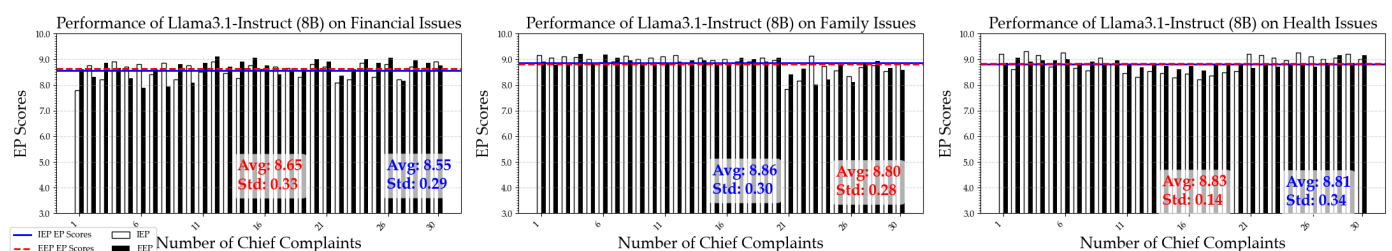


Figure 6. Llama 3.1-Instruct (8B) on the three scenario datasets.

6. Conclusions

The LLM-Agents framework presented in this study facilitates empathetic chatbot responses by leveraging distinct feature agents, including emotional, paraphrase, supportive, and attitudinal agents. The framework applies the IEP approach to decompose empathy into these features. Decomposing empathy into four features not only assists counseling professionals in understanding the impact of each aspect but also supports future adjustments and optimization of empathetic conversation quality. Experiment results with six models showed that utilizing these four empathy features enables the IEP-based design to

generate more empathetic dialogue outputs. Additionally, by calculating both the average and standard deviation of EP scores, the framework effectively analyzes the empathetic capabilities of counseling chatbots.

Author Contributions: Methodology, and writing—original draft preparation, Y.-R.C.; writing—review and editing, H.-Y.K.; supervision, K.-T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant NSTC 114-2223-E-007-011.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during the current study will be available in the near future: <https://github.com/jennifer-ctu/Empathetic-Chatbot> (accessed on 7 September 2025).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. The and potential of large language model based agents: A survey. *Sci. China Inf. Sci.* **2025**, *68*, 121101. [CrossRef]
2. Casas, J.; Spring, T.; Daher, K.; Mugellini, E.; Khaled, O.A.; Cudré-Mauroux, P. Enhancing conversational agents with empathic abilities. In Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA 2021), Tokyo, Japan, 14–17 September 2021; pp. 41–47. [CrossRef]
3. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; pp. 24824–24837.
4. Gosmar, D.; Dahl, D.A. Hallucination mitigation using agentic AI natural language-based frameworks. *arXiv* **2025**, arXiv:2501.13946. Available online: <https://arxiv.org/abs/2501.13946> (accessed on 7 September 2025).
5. Santos, J.M.; Shah, S.; Gupta, A.; Mann, A.; Vaz, A.; Caldwell, B.E.; Scholz, R.; Awad, P.; Allemanni, R.; Faust, D.; et al. Evaluating the Clinical Safety of LLMs in Response to High-Risk Mental Health Disclosures. *Res. Sq.* **2025**, preprint. [CrossRef]
6. Fu, J.; Ng, S.-K.; Jiang, Z.; Liu, P. GPTScore: Evaluate as You Desire. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 15–20 June 2024; pp. 6556–6576. Available online: <https://aclanthology.org/2024.naacl-long.365/> (accessed on 7 September 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.