# Can Large Language Models be Used to Provide Psychological Counselling? An Analysis of GPT-4-Generated Responses Using Role-play Dialogues

Michimasa Inaba, Mariko Ukiyo and Keiko Takamizo

Abstract Mental health care poses an increasingly serious challenge to modern societies. In this context, there has been a surge in research that utilizes information technologies to address mental health problems, including those aiming to develop counseling dialogue systems. However, there is a need for more evaluations of the performance of counseling dialogue systems that use large language models. For this study, we collected counseling dialogue data via role-playing scenarios involving expert counselors, and the utterances were annotated with the intentions of the counselors. To determine the feasibility of a dialogue system in real-world counseling scenarios, third-party counselors evaluated the appropriateness of responses from human counselors and those generated by GPT-4 in identical contexts in role-play dialogue data. Analysis of the evaluation results showed that the responses generated by GPT-4 were competitive with those of human counselors.

# 1 Introduction

Mental health care poses an increasingly serious challenge for modern societies. For example, in Japan, suicide is the leading cause of death among those aged 10 to 39 [10], and the World Health Organization has reported that suicide is a major cause of death among young people globally [12]. Consequently, text-based counselling, which refers to providing psychological support using instant messaging

Michimasa Inaba

The University of Electro-Communications, Chofu, Tokyo, Japan. e-mail: m-inaba@uec.ac.jp

Mariko Ukivo

iDEAR Human Support Service, Shibuya, Tokyo, Japan.

Keiko Takamizor

Japanese Organization of Mental Health and Educational Agencies, Shibuya, Tokyo, Japan. e-mail: takamizo@idear.co.jp

1

apps, has become increasingly prominent in recent years. Text-based counselling is more accessible than telecounselling and e-mail-based counseling, especially for the younger generation. However, the shortage of counselors proficient in text-based counseling is a major obstacle. Even those who are experienced in in-person and e-mail-based counseling, as well as telecounselling, encounter difficulties in providing text-based counseling without appropriate guidance and training. Moreover, there is a shortage of personnel who can provide such guidance.

Researchers are currently studying the uses of natural language processing for providing mental health support. Monitoring is an important aspect of mental health management, and the automatic detection of mental health disorders represents a particularly active area of current research [5, 1, 6, 11, 13].

In the field of dialogue systems, several systems aiming to improve mental health have been developed. For example, a system known as Woebot [4] facilitates cognitive behavioral therapy. There is also a system that provides emotional support [2, 9], and a system that responds by rewriting less empathetic utterances as more empathetic statements [14].

Recent developments with large language models (LLM) have demonstrated their adaptability to various tasks and domains. However, the performance of counselling dialogue systems that utilize LLMs has not been fully evaluated. The Chat-Counselor [8] system is the most similar to the system used in our study because this LLM-based model was constructed to answer clients' questions and requests as a counsellor. However, ChatCounselor only supports single-turn question-answering and does not support multiturn dialogue. Furthermore, in the study of ChatCounselor, the generated responses were evaluated automatically by GPT-4 and not by expert counsellors.

In this study, we constructed a counselling dialogue system using GPT-4, and professional counsellors evaluated the generated responses. To generate appropriate responses, we collected counselling dialogue data via role-playing scenarios involving expert counsellors, and the utterances were annotated with the intentions of the counsellors. To assess the feasibility of using a dialogue system in actual counselling situations, third-party counsellors evaluated the appropriateness of the responses given by the human counsellors and those generated by the GPT-4 during identical contexts in the role-play dialogue data.

#### 2 The collection of role play dialogue

In order to collect data for analysis, two counsellors participated in role-playing scenarios, with one counsellor playing the client and the other playing the counsellor. The dialogues were conducted in Japanese using the messaging application LINE, and data were collected for a total of six dialogues, one for each of the six themes shown in Table 1. Table 2 shows the number of words and utterances, and Table 3 shows a part of the collected role-play dialogues for Theme 1.

Table 1 List of role-play dialogue themes

6

1	(本番になると緊張してしまいます、どうしたらいいですか?(I get nervous
	at competitions. What should I do?)
2	パートナーのことは好きだけど束縛が強くてつらい (I love my partner, but
	it's hard because she is so restrictive.)
3	友だちとケンカしてしまいました、どうしたらいいですか?(I got into a
	fight with a friend. What should I do?)
4	勉強する気がありません、どうしたらいいですか? (I don't want to study.
	What should I do?)
5	男とか女とか決めつけられたくないです (I don't want to be categorized as
	male or female )

Table 2 The number of words and utterances from six role-play dialogues

	Utterances	Words	
Client	1,194	15,962	
Counsellor	1,140	31,689	
GPT-4 generated	816	43,833	

生きる意味ってなんだろう? (What is the meaning of life?)

**Table 3** Example of role-play dialogue and counsellors' key points and intent annotations for Theme 1. The abbreviation CLI stands for client, and COU stands for counsellor. The dialogues were conducted originally in Japanese and then translated into English by the authors.

Utterance		Key point	Intent
CLI:	I am a gymnast. Right before a competition, I get nervous and don't know what to do.		
COU:	I see.		To show that the counsellor is listening to the client.
COU:	Sometimes when you're nervous, you feel different than usual.		To convey that it is nothing special.
CLI:	By sorting out what things make you nervous, you may be able to find a way to cope.	0	
CLI:	For what? Not because it's a competition?		
COU:	For example, you may be nervous because everyone is watching you at a competition and you may wonder what they think of you,	long, send it once and	amples and to make it

# 3 Counsellor response generation by GPT-4

To analyze whether an LLM can successfully engage in counselling dialogues, we used the role-play dialogue data to prompt GPT-4 to generate utterances while assuming the role of a counsellor.

In LLM prompts, chain-of-thought (CoT) prompting [15] has been found to be effective at not only generating answers to questions and problems but also facilitating the process of thinking about them. To perform CoT prompting and generate higher-quality responses, we annotated the collected counsellor utterances with the response key points and the intent of the response (Table 3). For themes one through three, the annotations were performed by the counsellor-role speakers; for themes four through six, this was done by counsellors who did not participate in the role-play dialogue.

**Table 4** The prompt provided to GPT-4 for generation. To facilitate a process of thinking similar to chain-of-thought prompting, the counselor's key points and intentions annotated in the role-play data (the texts enclosed in "[]") were added before the counselor's utterances in the dialogue section. Sentences highlighted in green are GPT-4-generated intentions and responses. The prompt was originally written in Japanese and translated by the authors.

#### # Task

You provide text-based counseling. As a counselor, you should generate appropriate responses for your client and the intentions behind them. Intentions should be written in [] before the response.

#### # Counseling Guidelines

- · Show empathy to close distance with the client
- · Before expressing an opinion or asking a question, provide an agreeable response first
- · After offering an opinion, check how the client feels about it
- · Summarize the client's concerns and thoughts, and confirm that what the counselor understands is what the client wanted to talk about
- · Communicate a positive perspective that the client may not be aware of
- · Provide concrete examples to help the client visualize what is happening

#### # Counselling Dialogue

Counsellor: [Greetings] Good morning.

Client: Good morning

Counsellor: [Expresses gratitude for the client's decision to come for counselling. The client may have had the courage to come to you for counselling, and this is to put the client at ease.] Thank you for coming for counselling today.

Counsellor: [Asks broadly at first.] Please tell me what you would like to talk about today.

Client: OK

Client: Can I start now?

Counsellor: [Reassure the client that this is an environment where they could feel comfortable

to talk.] Sure. You can talk about whatever you want. Please feel free to do so.

The prompt we gave to GPT-4 for response generation is shown in Table 4. The prompt began with instructions telling GPT-4 to generate responses as a counsellor.

Moreover, counsellor guidelines that were developed under the supervision of professional counsellors were also provided to GPT-4. Next, we included the ongoing dialogue between the counsellor and the client in the prompts. The provided dialogue included all text from the start of the dialogue up until the previous client's utterance. In addition, the annotated key points and intentions were added before the counsellor's utterance in the dialogue, and the LLMs were instructed to generate them as well. Table 5 shows the generation examples, and the texts enclosed in "[]" are the generated intent.

We used GPT-4-0613 via OpenAI API, with a temperature of 0.0 and default values for the other parameters. Numerical details of the generated utterances are displayed in Table 2. As shown, the number of human counsellor utterances is higher than the number of utterances generated by GPT-4 because, in the role-play scenario, human speakers were able to send multiple consecutive messages. In contrast, GPT-4 generated one utterance at a time, as shown in Table 5.

**Table 5** Examples of intent and response generation results by GPT-4. The score represents the average of the three professional counsellors' evaluation scores for each generated response.

Role-play dialogue	GPT-4	Score
CLI: I don't have many friends, so she's the one who's been with me at school		
COU: I see.	[showing empathy] I see, your relationship with her is very important to you. [expresses understanding of the client's feelings] It's very painful that your relationship with her has turned out this way.	
CLI: Yes.		
it seems to me that you feel text mes- sages from her are bothersome and tha you don't feel you can talk about any- thing with her,		0.33
COU: So I asked you if she is what you call a "friend".		
CLI: I feel lonely being alone at school Don't you think so?		
ter, forcing yourself to be with some one you don't like very much, or being	[to empathize with the client] Sure, it's lonely to be alone at school, but it's important to create an environment where you feel comfortable. [to ascertain what the client would like to do] What would you like to do?	

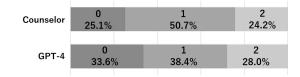
# 4 Analysis

# 4.1 Subjective evaluation by professional counsellors

To analyze the role-play dialogues and the GPT-4-generated utterances, third-party counsellors ranked both the utterances of the counsellors made during the role-play dialogues and the GPT-4-generated utterances. The utterances were rated according to a 3-point Likert scale from 0 (bad) to 2 (good), with three professional counsellors per dialogue. The counsellors also provided written explanations for their evaluations. A total of seven counsellors conducted evaluations. Table 5 shows examples of generated utterances and the average scores for each of the three counsellors' utterances.

We calculated Krippendorff's alpha for the dialogues of themes one through three (counselor utterances: 157; GPT-4 generated utterances: 124) in order to assess the agreement between the counselors' ratings of the utterances. The Krippendorff's alpha was 0.24, indicating a weak correlation.

The mean rating scores for the counsellors' utterances and GPT-4-generated utterances were 0.99 (variance: 0.49) and 0.94 (variance: 0.61), respectively. We also conducted a Mann-Whitney U-test at a significance level of 0.05 and found no significant differences. The above results indicate that the difference in response quality between counsellors and GPT-4 is minimal. Figure 1 shows the ratios of the ratings given to utterances by counsellors and GPT-4, respectively. As shown, a larger ratio of GPT-4 utterances was assigned a rating of 0 and 2 compared to the counsellor's utterances. More than half of the counsellors' utterances were given a score of 1; however, this was because the counsellors' utterances often included short utterances that were given a score of 1, such as "I see" and "Yes."



**Fig. 1** Ratio of rating scores assigned to utterances

The evaluation results confirm that there were individual differences in the scoring tendencies of the evaluators. Therefore, we analyzed whether the same evaluator gave higher scores to counsellors' utterances or GPT-4-generated utterances when the context was identical. In cases where the counsellor made multiple consecutive utterances before the client responded, the average rating for all utterances was used as the rating for the counsellor's utterance.

The results of the comparison are shown in Table 6. We found that more counsellors' utterances were rated as superior compared to GPT-4 utterances that were rated superior (34.8 % vs 30.5%). However, the difference was small, and 34.7%

of the utterances were rated as equal. We also confirmed that the quality of GPT-4 responses was very close to that of the human counsellors' responses.

Counsellor win	284 (34.8%)
GPT-4 win	249 (30.5%) 283 (34.7%)
Tie	283 (34.7%)

Table 6 Comparison of ratings for responses to the same context

We confirmed that LLM-generated responses were competitive with human responses despite having not used few-shot examples except for given annotated context in our prompt for GPT-4. Given that it is likely possible to further enhance the performance of GPT-4, we believe that there is potential for providing real-world counselling using an LLM-based system.

# 4.2 Case Study

When providing real-world counselling using a dialogue system, it is necessary to minimize inappropriate responses as much as possible. In this section, we analyze the responses generated by GPT-4 that were given low ratings.

The majority of responses with low ratings were phrases with inappropriate or unnatural phrasing and wording. An example of such an utterance is shown in Table 7. The counsellor who conducted the evaluation commented that the use of the word "interesting" could cause discomfort for the client because it could imply that the questions were being asked out of curiosity.

**Table 7** Example of a generated response with a low rating due to use of inappropriate wording (Theme 1)

Context	Councelor:	Councelor: Okay, I see.	
	Client:	I'm also fine when the person before me is performing.	
Generated	GPT-4:	That is interesting. So do you mean that you get more nervous right	
response		before your performance actually starts?	

There were also many responses where the GPT-4 response discusses the client's problem as if it were somebody else's problem. An example of such a case is shown in Table 8. In this example, two of the three evaluators commented that the response sounds insincere, and one commented that the expression "It looks difficult" should be avoided.

In particular, counsellors should take great care to avoid risky responses. A known problem with LLMs is that they can generate sentences with offensive or discriminatory content [16, 3]. The evaluators and the authors confirmed that no offensive or discriminatory statements were found among the GPT-4-generated ut-

**Table 8** Example of generated response with low rating because it treated the client's problem as if it were somebody else's problem (Theme 4)

Context	Counsellor:	If you don't mind, can you tell me when you feel like you are not studying as well as you would like? (You know what you have to do, but you don't have the motivation to do it. Or maybe you have a subject that you are not good at.)  I can't concentrate in class, and I usually fall asleep. That's why I can't do my homework I have to study for the entrance exam, but I don't
Generated response	GPT-4:	even know what to do.  I see. You won't be able to do your homework because you can't concentrate in class. It looks difficult that you don't know what to do with your exam studies.

terances. However, a small number of risky utterances were identified. An example of such an utterance is shown in Table 9. In response to this GPT-4-generated utterance, the evaluator wrote, "The response that kindness causes one to suffer oneself runs the risk of inculcating the value that one should not be kind."

**Table 9** Example of generated risky utterance (Theme 2)

Context	Counsellor: I see. You were thinking of her and putting her first. That is your kind-		
		ness, and she also made you that way.	
	Client:	Ah, yes. I've been told for some time that I'm too kind.	
Generated	GPT-4:	You are a kind person. But it is also true that your kindness has caused	
response		you to suffer.	

Although the number of risky responses identified in this study was small, when input prompts include offensive content, LLMs tend to generate offensive sentences in response [7]. In the role-play dialogues collected for this study, neither the counsellor nor GPT-4 generated responses included such content. In the real world, clients might provide responses with aggressive content. We plan to analyze such cases in future work.

#### **5** Conclusion

We collected and annotated role-play counseling dialogue data, and professional counselors evaluated the appropriateness of GPT-4-generated responses. Analysis of the results demonstrated that the GPT-4-generated responses were competitive with those of human counselor responses. Furthermore, among the responses that received low ratings, no aggressive, discriminatory, or high-risk responses were identified.

In this study, we generated and evaluated responses based on the context of roleplay dialogues. No experiments have demonstrated a dialogue system with a fully automatic counseling dialogue from start to finish. In future work, we plan to evaluate whether this system can provide fully automated, comprehensive counseling services.

#### References

- Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. NPJ digital medicine, Vol. 3, No. 1, p. 43, 2020.
- Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. PAL: Personaaugmented emotional support conversation generation. In Findings of the Association for Computational Linguistics: ACL 2023, pp. 535–554, 2023.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models, 2023.
- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, Vol. 4, No. 2, p. e7785, 2017.
- George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. Scientific reports, Vol. 7, No. 1, p. 45141, 2017.
- Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. *Complexity*, Vol. 2018, , 2018.
- Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. Certifying llm safety against adversarial prompting, 2023.
- 8. June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chatcounselor: A large language models for mental health support, 2023.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 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), pp. 3469–3483, 2021.
- 10. Ministry of Health, Labour and Welfare of Japan. White paper on suicide prevention, 2020.
- 11. Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings* of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8446–8459, 2022.
- World Health Organization. World mental health report: Transforming mental health for all, 2022.
- Tulika Saha, Saichethan Miriyala Reddy, Sriparna Saha, and Pushpak Bhattacharyya. Mental health disorder identification from motivational conversations. *IEEE Transactions on Computational Social Systems*, 2022.
- Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. WWW '21, p. 194–205. Association for Computing Machinery, 2021.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, Vol. 35, pp. 24824–24837, 2022.
- Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity, 2023.