Exploring User Feedback: A Thematic and Sentiment Analysis of User Interactions with LLM-Based Dialogue Robots

Muhammad Yeza Baihaqi[1,2] Angel García Contreras [2] Seiya Kawano[2,1] Koichiro Yoshino[3,2,1]
[1]Nara Institute of Science and Technology [2]Guardian Robot Project, RIKEN
[3]Institute of Science Tokyo
muhammad_yeza.baihaqi.lx2@naist.ac.jp
{angel.garciacontreras, seiya.kawano}@riken.jp
koichiro@c.titech.ac.jp

Abstract

Recent advancements in large language models (LLMs) have significantly improved dialogue agents, enabling them to generate context-aware, human-like responses. While quantitative evaluations effectively compare performance based on predefined metrics, they may fail to capture nuanced user experiences, such as memorable exchanges or unexpected opinions, which are crucial for refining the system. To address this issue, we conducted thematic and sentiment analysis by collecting participant feedback through dialogue experiments. Specifically, we assessed GPT-3.5-Turbo and GPT-4o as dialogue models for dialogue robots. Thematic analysis allowed us to identify recurring patterns in user experiences, while sentiment analysis helped gauge the emotional tone of those interactions. Our experimental results provided rich insights in the form of themes and sub-themes, such as perceptions of knowledge depth and mistake correction. Sentiment analysis complemented these findings, showing that GPT-4o received a positive impression, while GPT-3.5-Turbo garnered mostly negative feedback.

1 Introduction

The recent advancements in large language models (LLMs) have ushered in a new era for dialogue agents [1]. With LLM-based dialogue models, dialogue agents can generate highly coherent, context-aware, and human-like responses, significantly enhancing their conversational capabilities [2, 3, 4]. Nowadays, they are being utilized in a variety of scenarios, such as counseling [5], pharmaceuticals [6], and small talk rapport agents [7].

To evaluate the performance of these dialogue agents, human evaluation through quantitative analysis is commonly employed. Our recent research utilized Likert scales and pairwise comparison questionnaires to assess the rapport-building capabilities of dialogue agents through quantitative analysis [8]. These methods quantify agents' performance in predefined attributes, such as user satisfaction and engagement, providing insights into specific aspects of agent performance.

Quantitative analysis, while effective for identifying trends and comparing performance, has notable limitations in capturing the subtleties of the studied aspects [9]. This approach is often confined to predefined variables, making it inappropriate for capturing unpredictable events that may arise during interactions [9, 10, 11]. For example, while numerical scores may show that one agent is perceived as more natural than another, they fail to uncover the underlying reasons—whether due to the joy it elicited, its ability to maintain a logical flow, or other unforeseen factors. Additionally, a naturalness score alone cannot determine whether users felt the agent was close to human-like or merely an incremental improvement. Without collecting participant feedback, these critical nuances remain unexplored, limiting our ability to adapt to unexpected outcomes and refine the system accordingly.

To address these limitations, qualitative analysis called thematic analysis offers a complementary approach by exploring user feedback in greater depth. This method involves analyzing qualitative data, such as user comments, to identify recurring themes and patterns that more holistically describe the experiences and emotions of users during their interactions [12]. These user comments often include

unique details, such as specific memorable exchanges, unexpected user opinions, or the agent's handling of a particular situation, which are difficult to capture by quantitative metrics alone [13]. Ultimately, this approach provides rich and varied data that transcends expected variables [9].

Thematic analysis is also commonly combined with sentiment analysis. Sentiment analysis involves analyzing people's opinions and sentiments toward entities [14]. Combining sentiment analysis with thematic analysis allows researchers to gain deeper insights into not only the overarching themes present in the data but also the emotional tone associated with those themes [15]. This combination enables a more nuanced understanding of how participants feel about specific themes, helping to identify patterns of sentiment within different thematic categories [16].

Given these advantages, this study aims to leverage thematic and sentiment analyses to gain a richer understanding of user experiences with dialogue robots. Specifically, this experiment compares GPT-4o and GPT-3.5-Turbo as dialogue models for dialogue robots. We conducted dialogue experiments, gathering detailed participant feedback to uncover nuanced user experiences and provide unique insights into the differences between the two dialogue robots.

2 Methodology

2.1 Participants

The study involved 20 participants, equally divided between 10 males and 10 females, with an average age of 23.35 years (SD = 1.02). None had prior experience with robot interactions. Informed consent for data usage was obtained from all participants before the experiment.

2.2 Dialogue systems and robots

In this research, we specifically utilized the GPT-3.5-Turbo and GPT-4o models. Both large language models employed a free-form approach to prompt rapport-building dialogue systems [7]. The dialogue strategy focused on integrating rapport-building utterances into small talk. These utterances, derived from proven human-to-human interactions, included techniques such as praising, encouragement, and recommendations, among others, to foster rapport. Additionally, the system employed two types of questions—short questions and open-ended questions—to ensure conversational continuity until 28 turns.

Figure 1 CommU robot.

The LLMs were integrated into dialogue robots named CommU, as shown in Figure 1. Participants communicated with the robots through voice interaction. To enable this, we utilized Julius-based automatic speech recognition (ASR) [17] to capture participants' voices and employed a text-to-speech (TTS) system[1)] to generate the robot's voice. Dialogue robots powered by GPT-3.5-Turbo were referred to as BotA, while those using GPT-4o were named BotO.

2.3 Experimental procedure

In this experiment, we used a counterbalanced design. Each participant engaged in a small talk with both BotA and BotO. After interacting with both robots, participants were asked to provide their experiences in the form of short or long comments for each robot.

2.4 Analyses

2.4.1 Thematic analysis

Thematic analysis was conducted by analyzing the participants' comments for each robot. First, user comments were coded based on their characteristics. Short comments, typically consisting of a single sentence, could be directly classified into a theme and sub-theme. On the other hand, longer comments required a coding process before classification. If a user made a long comment like, "*The reactions were natural and similar to those of humans. It was easy to have a conversation because he introduced me to different topics and asked me questions to dig deeper*," we would first break the comment into shorter sentences and then classify them into relevant themes and sub-themes.

1) https://pypi.org/project/gTTS/

These comments were categorized into specific themes and sub-themes through an iterative process. The themes and sub-themes emerged directly from participants' comments, rather than being shaped by predetermined frameworks or theories [18]. This approach allows us to capture rich data reflecting the perspectives of the participants. Additionally, in this research, we specifically focused on assessing both LLMs as the dialogue models of a dialogue robot, and comments outside these domains—such as those related to the robot's appearance—were discarded to maintain focus on relevant aspects.

2.4.2 Sentiment analysis

Sentiment analysis was performed to classify the sentiment of each coded comment as positive or negative. To do this, the comments were initially fed into ChatGPT, which provided an automatic classification based on the detected emotional tone. Afterward, the researchers manually verified the classifications to ensure accuracy and consistency.

3 Results

3.1 Thematic findings

After carefully going through each comment iteratively, we decided on three themes: Conversation Behavior (CB), Conversation Content (CC), and Conversation Flow (CF). For BotA, we had 19 comments, and for BotO, we had 27 comments. The example comments are shown in Appendices A1 and A2. The comparison of BotA and BotO, based on our thematic analysis, is shown in Table 1.

3.1.1 Conversation behavior

For BotA, the theme of CB emerged prominently, with a focus on Friendliness and Human-likeness. Participants expressed dissatisfaction with the bot's lack of friendliness, noting that it did not feel like conversing with a friend and lacked personal opinions. Additionally, BotA was perceived as less natural compared to BotO, indicating shortcomings in its human-like qualities.

BotB exhibited stronger CB, particularly in Communication Skills and Human-likeness. While BotA and BotO also used rapport-building strategies, participants praised BotB's natural joke delivery, topic guidance, and conversational style, which made it feel less robotic. Many noted that its reactions felt so natural it hardly seemed like a machine, with slight inconsistencies adding to its conversational authenticity. BotB also excelled in Mistake Correction, demonstrating impressive comprehension and adaptability in handling errors.

3.1.2 Conversation content

For BotA, the theme of CC emerged with sub-themes of Enjoyment and Knowledge Depth. Participants mentioned fun discussions and surprising topics like "friendship," but opinions on knowledge depth were mixed. While some appreciated recommendations like summer dishes, others found the content lacking depth.

In contrast, BotO exhibited stronger performance in CC, with sub-themes of Enjoyment, Knowledge Depth, and Satisfaction. Participants frequently noted the bot's humorous and lively storytelling, which made the conversations enjoyable. The bot also displayed greater knowledge depth, surprising users with unique and detailed information, such as lesser-known travel destinations and specific suggestions tailored to the conversation. These qualities contributed to a higher level of satisfaction.

3.1.3 Conversation flow

For BotA, the theme of CF emerged with sub-themes of Conversation Ending and Effort in Topic Transition. Participants expressed confusion and awkwardness about how conversations ended, with repeated goodbyes and a lack of new topics leading to stagnation. Additionally, users highlighted difficulties in transitioning between topics, as the robot often required them to introduce new topics or simply echoed their statements. This lack of proactive engagement made conversations challenging to navigate.

In comparison, BotO demonstrated mixed performance in CF, also encompassing the sub-themes of Conversation Ending and Effort in Topic Transition. While some participants found the endings awkward, they noted that BotO showed greater effort in maintaining topic transitions. The bot's use of relatable examples and consistent questioning made it easier for users to continue conversations and feel engaged. However, occasional abrupt topic changes and unclear progression were noted as areas for improvement.

3.2 Sentiment in comments

BotA shows predominantly negative sentiment in conversation behavior and flow (100%) and mixed sentiment

Table 1 Comparison of BotA and BotO based on Thematic Analysis.

Theme	Sub-theme	BotA (GPT-3.5-Turbo)	BotO (GPT-4o)
CB	Friendliness	BotA lacked friendliness, with responses feeling neutral and impersonal.	Conversation with BotO feels like interacting with a friend.
	Human-likeness	BotA is considered less natural than BotO.	BotO was perceived as natural, with users feeling that the topic guidance and reactions were human-like, to the point where they did not realize they were interacting with a robot.
	Communication skill	Not explicitly mentioned	Users perceived the communication skills of BotO were improved compared to BotA.
	Mistake correction	Not explicitly mentioned	BotO responded appropriately to corrections, showing understanding and adaptability when users identified mistakes.
CC	Enjoyment	Users found the conversations fun at times, with occasional surprises.	Users consistently described conversations as fun, engaging, interactive, and entertaining with humorous remarks and lively exchanges.
	Knowledge depth	BotA provided basic information, such as seasonal dish recommendations. Content lacked depth and could feel generic.	BotO demonstrated deeper knowledge by offering unique insights, detailed suggestions, and relevant keywords, which were useful for broadening the conversation scope.
	Satisfaction	Not explicitly mentioned	Users explicitly stated being more satisfied with BotO conversations.
CF	Effort in topic transition	Users reported challenges with topic transitions, citing BotA's passive responses, lack of questions, and repetitive agreement statements, which made it difficult and confusing to move the conversation forward.	It was noted that continuing the conversation with BotO was much easier than with BotA due to its content and behavior. However, a user noted that the topic changes were abrupt.
	Ending the conversation	Towards the end of the conversation, it felt awkward and stagnant, with no new topics to discuss, repeated goodbyes, and a sense of confusion about how it concluded.	Towards the end of the conversation, there were still two reports noting that it got stuck and involved repeated goodbyes, though this occurred less frequently than with BotA.

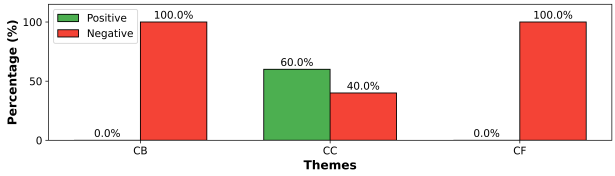

Figure 2 Sentiment analysis of BotA.

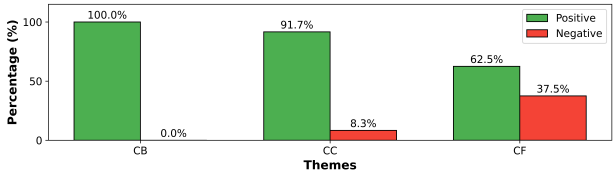

Figure 3 Sentiment analysis of BotO.

in content (60% positive). In contrast, BotO achieves positive sentiment in behavior (100%) and content (91.7%), with moderate results in flow (62.5% positive). These results suggest that BotO provides a more engaging and satisfactory user experience compared to BotA.

4 Conclusion

In this research, we employed both thematic and sentiment analyses to examine participants' feedback on two LLMs, GPT-3.5-Turbo and GPT-4o, as dialogue models for a dialogue robot. Our thematic analysis uncovered important themes and sub-themes related to user experience, such as perceived friendliness and the effort involved in topic transitions. These insights provided a deeper understanding of how participants engaged with the models, beyond what could be captured through quantitative methods alone. It was found that GPT-4o outperformed GPT-3.5-Turbo in nearly all aspects. While both models received positive feedback regarding enjoyment, issues with conversation flow persisted for both. Sentiment analysis revealed that GPT-4o generally receiving positive sentiment and GPT-3.5-Turbo receiving more negative feedback. Future research could replicate this study with a larger and more diverse sample, incorporating semi-structured interviews to gather more detailed feedback.

Acknowledgement

A part of this work is supported by JSPS KAKEN-HI Grant Number 23K24910 and 23K19984.

References

[1] Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. A survey of the evolution of language model-based dialogue systems, 2023.

[2] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models, 2024.

[3] Varun Nair, Elliot Schumacher, Geoffrey Tso, and Anitha Kannan. DERA: Enhancing large language model completions with dialog-enabled resolving agents. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Danielle Bitterman, editors, **Proceedings of the 6th Clinical Natural Language Processing Workshop**, pp. 122–161, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] Z. Ma, Y. Mei, and Z. Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In **Proceedings of the AMIA Annual Symposium**, pp. 1105–1114, January 2024.

[5] Michimasa Inaba, Mariko Ukiyo, and Keiko Takamizo. Can large language models be used to provide psychological counselling? an analysis of gpt-4-generated responses using role-play dialogues. In **Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS)**, pp. 1–9, March 4–6 2024.

[6] Vania Amanda Samor, Muhammad Yeza Baihaqi, Edmun Halawa, Luh Rai Maduretno Asvinigita, Sarah Nabila Hakim, and Mela Septi Rofika. Evaluating llms as pharmaceutical care decision support tools across multiple case scenarios. In **Proceedings of the International Conference on Medical Science and Health (ICOMESH 2024)**, pp. 273–282. Atlantis Press, 2024.

[7] Muhammad Yeza Baihaqi, Angel Garcia Contreras, Seiya Kawano, and Koichiro Yoshino. Rapport-driven virtual agent: Rapport building dialogue strategy for improving user experience at first meeting. In **Interspeech 2024**, pp. 4059–4063, 2024.

[8] Muhammad Yeza Baihaqi, Angel García Contreras, Seiya Kawano, and Koichiro Yoshino. Comparing likert scale and pairwise comparison for human evaluation in rapport-building dialogue systems. Technical Report 43, Nara Institute of Science and Technology / Guardian Robot Project RIKEN / Institute of Science Tokyo, December 5 2024.

[9] Daniel Eyisi. The usefulness of qualitative and quantitative approaches and methods in researching problem-solving ability in science education curriculum. **Journal of Education and Practice**, Vol. 7, No. 15, pp. 91–100, 2016.

[10] M. Denscombe. **The Good Research for Small-Scale Social Research Project**. Open University Press, Philadelphia, 1998.

[11] J. W. Creswell. **Research Design: Qualitative, Quantitative, and Mixed Methods Approaches**. SAGE Publications, London, 3rd edition, 2009.

[12] T. Vandemeulebroucke, B. D. de Casterlé, and C. Gastmans. How do older adults experience and perceive socially assistive robots in aged care: a systematic review of qualitative evidence. **Aging & Mental Health**, Vol. 22, No. 2, pp. 149–167, February 2018. Epub 2017 Feb 9.

[13] B. Zhao, J. Lam, H. M. Hollandsworth, A. M. Lee, N. E. Lopez, B. Abbadessa, S. Eisenstein, B. C. Cosman, S. L. Ramamoorthy, and L. A. Parry. General surgery training in the era of robotic surgery: a qualitative analysis of perceptions from resident and attending surgeons. **Surgical Endoscopy**, Vol. 34, No. 4, pp. 1712–1721, April 2020. Epub 2019 Jul 8.

[14] Basant Agarwal and Namita Mittal. Optimal feature selection for sentiment analysis. In Alexander Gelbukh, editor, **Computational Linguistics and Intelligent Text Processing**, pp. 13–24, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[15] E. Ainley, C. Witwicki, A. Tallett, and C. Graham. Using twitter comments to understand people's experiences of uk health care during the covid-19 pandemic: Thematic and sentiment analysis. **Journal of Medical Internet Research**, Vol. 23, No. 10, pp. 1–14, October 25 2021.

[16] E. L. Funnell, B. Spadaro, N. Martin-Key, T. Metcalfe, and S. Bahn. mhealth solutions for mental health screening and diagnosis: A review of app user perspectives using sentiment and thematic analysis. **Frontiers in Psychiatry**, Vol. 13, pp. 1–17, April 27 2022.

[17] Akinobu Lee and Tatsuya Kawahara. julius-speech/julius: Release 4.5, January 2019.

[18] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. **Qualitative Research in Psychology**, Vol. 3, No. 2, pp. 77–101, 2006.

A Appendix

A.1 Example comments for BotA

- **Conversation behavior: Friendliness**
  "I felt like I was not talking to a friend because it (robot) did not have a personal opinion." [Participant 2, age range: 23, female]
- **Conversation flow: Ending the conversation**
  "It felt awkward to have to say goodbye and hello so many times at the end of a conversation." [Participant 9, age: 24, female]
- **Conversation content: Knowledge depth**
  "It (robot) touched on various topics, but I felt the content lacked depth." [Participant 5, age: 23, male]

A.2 Example comments for BotO

- **Conversation behavior: Mistake correction**
  "When I answered the question, I said something wrong but later corrected my mistake. I was surprised at how well he understood." [Participant 19, age: 23, male]
- **Conversation flow: Effort in topic transition**
  "It was difficult to follow the progression of the conversation, as the topic suddenly changed, making it hard to continue." [Participant 12, age: 22, male]
- **Conversation content: Knowledge depth**
  "I was surprised when it (robot) started talking about information about travel destinations that were not well-known." [Participant 15, age: 24, male]