Empathy vs Neutrality: Designing and Evaluating a Natural Chatbot for the Healthcare Domain

Cristina Reguera-Gómez
TNO
Utrecht University
c.regueragomez@uu.nl

Denis Paperno
Utrecht University
d.paperno@uu.nl

Maaike H.T. de Boer
TNO
maaike.deboer@tno.nl

Abstract

As lifestyle-related diseases rise due to unhealthy habits such as smoking, poor diet, lack of exercise, and alcohol consumption, the role of Conversational AI in healthcare is increasingly significant. This study provides an empirical study on the design and evaluation of a natural and intuitive healthcare chatbot, specifically focusing on the impact of empathetic responses on user experience regarding lifestyle changes. Findings reveal a strong preference for the empathetic chatbot, with results showing statistical significance (p <0.001), highlighting the importance of empathy in enhancing user interaction with healthcare chatbots.

1 Introduction

In our contemporary healthcare situation, lifestyle-related diseases are increasing, primarily influenced by unhealthy habits such as smoking, poor diet, lack of exercise, and alcohol consumption (Balwan and Kour, 2021). Simultaneously, conversational AI, or chatbots, have gained popularity and emerged as powerful tools, particularly in the healthcare sector (Amiri and Karahanna, 2022).

Nevertheless, the use of conversational agents in the healthcare domain is not too widespread, especially when compared to other industries such as travel and hospitality (Laranjo et al., 2018). Furthermore, very little is known about how the linguistic design of a medical conversational agent can impact the users' likelihood to employ it for their healthcare queries (Shan et al., 2022).

This paper focuses on the design and evaluate a natural and intuitive chatbot for the healthcare domain, including an empirical analysis of the results. More specifically, we investigate how the use of empathy in generated messages can affect user experience during queries about lifestyle changes, hence influencing the likelihood to incorporate a healthcare conversational agent in their daily lives (de Boer et al., 2023). The two primary contributions of this study are:

1. To provide insight in the impact of empathetic versus neutral tones in messages in a LLM based chatbot.

2. To understand user expectations in human-computer interactions - using chatbots - in the healthcare domain, especially on lifestyle changes.

2 Related Work

2.1 Empathy and Language in Human-Computer Interaction (HCI)

Empathy plays a crucial role in making HCI more natural and intuitive. This paper draws on concepts of cognitive and affective empathy in human interaction.

Empathy is generally divided into two types: cognitive and affective empathy. Cognitive empathy is the ability to understand another person's emotional state without necessarily sharing it. Reniers et al. (2011) describe cognitive empathy as constructing a mental model of another's emotions. For example, someone with strong cognitive empathy can understand a friend's distress over a failure and offer appropriate advice. Cognitive empathy facilitates communication by enabling deeper understanding of others' experiences.

In contrast, affective empathy involves an emotional response to another's feelings. Affective empathy allows individuals to emotionally connect with others by vicariously experiencing their emotions. For instance, when a friend celebrates an achievement, a person with affective empathy would also feel joy. This type of empathy is essen-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 508–517
March 3-4, 2025 ©2025 University of Tartu Library

tial for providing emotional support and fostering deeper connections.

Empathy is fundamental in social cognition (Iacoboni, 2005), allowing individuals to share experiences and goals. While empathy in humans involves complex cognitive and emotional mechanisms, chatbots can replicate empathetic communication by imitating patterns of human interaction. In practice, empathetic language in chatbots focuses on word choices that acknowledge the user's emotional state, effectively simulating empathy through language.

Human empathy, as standardly defined (Cuff et al., 2016), involves complex mental and emotional processes that chatbots do not possess. Instead, when discussing empathy in chatbots, we refer to their ability to produce responses that mimic human empathetic behaviours. Henceforth, a chatbot can be considered empathetic if its responses create the illusion of understanding and validating the user's feelings, even though it lacks real emotional experience.

2.2 Language Choices in Empathetic Communication

Empathetic communication in chatbots is achieved not only through understanding emotions but also through specific linguistic choices. Research by Yaden et al. (2023) identifies words associated with empathy, showing how language can create a sense of emotional support and connection. For example, the use of personal pronouns such as "I" and "you" helps create a more direct and personal interaction. Similarly, adjectives like "good" and "happy" convey positive emotional states, while verbs like "hope" and "need" can express concern or reassurance.

In addition to word choices, certain phrases play an essential role in empathetic communication. Lapointe (2014) found that common phrases like "I know" and "I understand" are often used to validate the user's feelings, while phrases like "it is" and "you are" are used to acknowledge the situation. These phrases help build emotional connection and foster a sense of understanding between the speaker and listener, which is crucial in emotionally sensitive interactions.

2.3 Research on Empathy in Chatbots

Research has increasingly focused on how empathetic language in chatbots can enhance user experience. Liu and Sundar (2018) explored whether chatbots should offer both informational and emotional support when advising on personal issues. Their findings show that users generally prefer empathetic expressions over neutral advice, even when delivered by a chatbot, particularly when users are skeptical of machines' ability to show empathy.

Casas et al. (2021) further investigated empathetic chatbots by developing a system that generates emotionally attuned responses. Their chatbot outperformed both a standard chatbot and even some human responses in terms of perceived empathy. These studies demonstrate that empathetic language significantly improves user satisfaction with chatbots.

In the healthcare domain, the BabyTalk project (Mahamood and Reiter, 2011) examined parental preferences for emotionally sensitive medical reports about babies in neonatal care. Parents overwhelmingly preferred emotionally supportive, or affective, language over neutral descriptions. This shows that empathetic language is not only valued but essential in high-stress environments.

3 Conversational Agent Design

The decision to use a LLM-powered chatbot was driven by the need for a system capable of understanding and generating natural language with a high degree of fluency and contextual awareness. Unlike traditional rule-based or retrieval-based chatbots, which rely on predefined scripts or a database of responses, an LLM-powered chatbot can generate nuanced, contextually appropriate responses based on the specific needs of the user at any given moment.

One of the primary advantages of LLMs is their ability to process complex language inputs, making them well-suited for conversations that require deep contextual understanding, such as those in the healthcare domain. Given the nature of healthcare queries, which often involve detailed and sensitive information, it was essential to implement a system that could handle such complexities with a high degree of accuracy and flexibility.

The main objective of the chatbot is generating responses to user queries in a manner that is both informative and aligned with the specific version (empathetic or neutral) being tested.

The implementation of both the empathetic and neutral version is the same, except for the specific prompt used. In our first experiment, we evaluate different LLMs to decide on the most suitable for our task.

The implementation of the chatbots involved using the AutoTokenizer from Hugging Face to preprocess and tokenise input data, ensuring compatibility with the model and efficient handling of user queries. The chatbot's LLM was run locally using a GPU cluster, which was crucial for managing the computational demands of real-time text generation during user interactions. The web component of the chatbot was built using Flask, a lightweight web framework for Python, chosen for its simplicity and effectiveness in developing web-based applications.

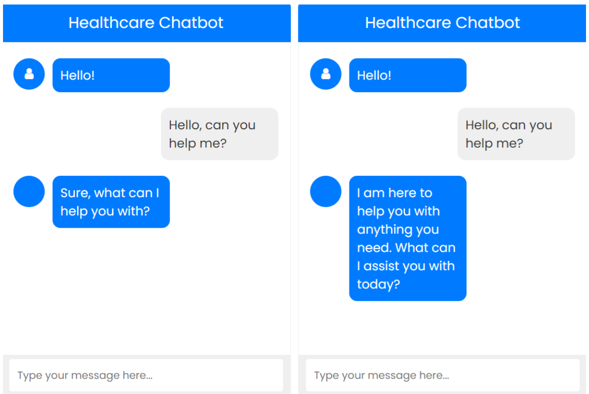
3.1 Empathetic Chatbot Design

The empathetic version of the chatbot was designed with a specific focus on enhancing user experience through emotionally supportive communication. This required a detailed approach to ensure that the generated responses not only conveyed the necessary information but did so in a manner that validated and supported the user's feelings. To implement empathy in the generated responses, the empathetic chatbot was programmed to follow a predefined prompt of empathetic communication, which diverges from that of the neutral one, and that was designed to shape its tone and language. The prompt explicitly instructs the model to generate responses that include empathetic expressions, focusing on word choices that reassure and validate the user's experiences. This approach ensures that the chatbot's interactions are not only informative but also emotionally supportive, thereby enhancing the overall user experience:

- **Neutral prompt**: "You are a chatbot who provides advice about lifestyle changes."

- **Empathetic prompt**: "You are a friendly chatbot who provides advice about lifestyle changes. Your responses must be empathetic. A response is considered empathetic if it shows: 1) Comprehension towards the feelings of the other (i.e. 'I understand that you are concerned about your health.'), and 2) Engagement in the feelings of the other (i.e. 'I feel so happy that you have decided to live a healthier lifestyle.'). Remember, your generated advice should contain word choices that reassure and validate other people's experiences, according to the definition of an empathetic response."

The empathetic prompt is characterised by two key elements in empathetic language that align with the literature, as they were mentioned in the previous section: comprehension of feelings (cognitive empathy) and engagement with feelings (affective empathy). The chatbot acknowledges and understands the user's emotions, providing responses that appeal to the user's emotional state. For example, it might say, "I understand that you are concerned about your health." to validate the user's concerns. Furthermore, the chatbot expresses positive reinforcement and encouragement, aiming to motivate the user. For instance, "I feel so happy that you have decided to live a healthier lifestyle!" is used to engage with and uplift the user. This prompt differs from that of the neutral chatbot, which only was instructed to provide advice about lifestyle changes, without any remark about the tone employed (see Figure 1). The prompts were chosen to meet the generative task requirements and to align with existing literature. They were also refined to ensure that the LLM could accurately understand the type of message it was asked to generate.



Hello!

Hello, can you help me?

Sure, what can I help you with?

Hello!

Hello, can you help me?

I am here to help you with anything you need. What can I assist you with today?

Figure 1: User interface and greetings generated by the neutral chatbot (left) and empathetic chatbot (right).

4 Experiment I: LLM Evaluation

The first experiment consisted of an evaluation of the responses generated by different LLMs, where the best-performing LLM was used as the basis of the chatbot in the user experiment (experiment 2).

4.1 Dataset

In order to do so, we asked the models to generate answers to questions obtained from the MASH-QA dataset (Zhu et al., 2020). This dataset was chosen because it is composed of consumer healthcare queries sourced from the popular health website WebMD, which features a wide range of articles covering various consumer healthcare topics. The answers to these queries are drawn from sentences or paragraphs within the articles related to the specific healthcare condition. These responses are curated by healthcare experts to ensure they accurately address the questions. We selected 100 questions, divided equally according to the following topics: exercise, food, smoking and alcohol. These topics are the same we used during the user experiment, since they are the related to the most common causes of lifestyle diseases.

4.2 Models

We chose four LLMs, two of them being domain-specific—MedAlpaca (Han et al., 2023) and Meditron (Chen et al., 2023)—and the other two being general—GPT-4 (OpenAI, 2024) and Llama 3 (Meta, 2024). The motivation behind this choice is that it is crucial to experiment with a diverse set of LLMs, due to the lack of agreement in the literature over the superior performance of general or domain-specific models for medical tasks (Zhou et al., 2024; Nori et al., 2023). Henceforth, the two most suiting domain-specific LLMs were chosen, along with two general ones: one that had yielded good results for medical tasks (GPT-4), and a powerful, open-source one (Llama 3).

4.3 Evaluation

The evaluation was performed with G-Eval (Liu et al., 2023), a state-of-the-art NLG evaluation framework that uses a chain-of-thought (CoT) and a form-filling paradigm to assess the quality of texts generated by LLMs with GPT-4. The primary benefit of this evaluation framework is that it achieves a higher correlation (0.588) with human judgments compared to conventional metrics and previously established LLM-based evaluators, such as BLEU or ROUGE.

For this study, we tested G-Eval with the following metrics:

1. Fluency: "the quality of the answer in terms of grammar, spelling, punctuation, word choice, and sentence structure" (Fabbri et al., 2021, as cited in (Liu et al., 2023)).
2. Coherence: "the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby 'the answer should be well-structured and well-organized. The answer should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic'" (Fabbri et al., 2021, as cited in (Liu et al., 2023)).
3. Groundedness: "the use of a fact in the answer, given the fact that this answer is conditioned by it" (Mehri & Eskenazi, 2020, as cited in (Zhong et al., 2022)).
4. Naturalness: "the quality of the answer in terms of being like something a person would naturally say" (Mehri & Eskenazi, 2020, as cited in (Zhong et al., 2022)).

These metrics were chosen because they encompass linguistic aspects related to human-likeness and user experience, so the scores associated with them can shed light on which models perform best on these aspects. In other words, this evaluation gives insights into how each LLM performs on the task of advice about lifestyle changes, from a linguistic point of view.

	GPT-4	Llama 3	MedAlpaca	Meditron
Fluency	**0.865**	0.864	0.844	0.846
Coherence	**0.753**	0.732	0.726	0.694
Groundedness	0.883	0.879	0.851	**0.894**
Naturalness	0.806	0.787	**0.826**	0.818
Avg. scores	**0.827**	0.816	0.812	0.813

Table 1: Results of the LLM evaluation.

4.4 Results and Discussion

As shown in Table 1, all the models had similar average scores across every metric, within the range 0.812 - 0.827, and with GPT-4 giving the highest score. Nevertheless, the results of the one-way ANOVA indicated that none of the differences between models in any metric were statistically significant ($p > 0.05$).

GPT-4 outperformed the other models in fluency (0.865) and coherence (0.753), which illustrates its linguistic abilities to generate dialogues.

The model's scores in fluency and coherence indicate its advanced linguistic capabilities, which are crucial for dialogue. Its fluency score (0.865) reflects the model's ability to produce smooth, easily readable text. GPT-4's coherence score (0.753) also surpasses the other models, suggesting that GPT-4 maintains logical consistency and context better throughout its responses. However, other individual metrics show slightly different results.

Meditron, one of the domain-specific models, received the highest score on groundedness (0.894), where the generated answers where compared about those from the golden standard in the MASH-QA dataset. Groundedness measures the factual accuracy and alignment of generated answers with a predefined gold standard, in this case, the MASH-QA dataset. Meditron's domain-specific training likely enhances its ability to produce accurate, relevant information within its specialised area. This specialisation illustrates the trade-off between general linguistic capabilities and domain-specific accuracy. While other models surpass Meditron in metrics concerning general dialogue quality, Meditron provides more precise and reliable information in the medical field.

The most surprising aspect of the data is in the results of naturalness, where MedAlpaca outperformed the other models (0.826). Naturalness evaluates how human-like the generated responses are, which is critical for creating engaging interactions. Despite MedAlpaca not leading in overall average scores or in fluency and coherence, its top performance in naturalness suggests that its generated messages are more intuitively aligned with human conversational patterns. Since naturalness was the most important metric, due to its relation to human-likeness, MedAlpaca was the chosen model to embed in the conversational agent of the main experiment.

5 Experiment II: User Experiment

We further compared the user experience with the neutral and empathetic conversational agents based on MedAlpaca.

5.1 Procedure

The experiment consisted of randomised controlled trials followed by cross-sectional surveys. A total of 25 participants were recruited, all of whom had completed university-level education. Of these participants, 68% identified as women, and 32% identified as men. In terms of age distribution, 68% were between 25 and 34 years old, while 12% were either 18 to 24 years old or 35 to 44 years old. Additionally, 4% were aged 45 to 54 years, or 55 to 64 years. The participants did not necessarily search for lifestyle change. They interacted with the chatbot remotely and were instructed to complete the experiment in a quiet environment. The independent variables included factors such as the participant demographics, empathy condition and scenario.

An initial questionnaire was used to gather information on personal information such as age and gender, and a 5-point Likert scale questionnaire on the following topics: frequency of use with chatbots, feelings towards chatbot use, and feelings towards chatbot use in healthcare.

A within-subject design was used, where the same participant tested all conditions. During the experiment, they interacted with a chatbot and asked for lifestyle advice according to the following scenarios they enacted: eating healthier, exercising more, quitting smoking and reducing alcohol intake. After each of those four scenarios, they filled in a questionnaire.

5.2 Materials

The questionnaire used to test the participants' interaction was an adapted version of the Chatbot Usability Questionnaire (CUQ) (Holmes et al., 2019). The CUQ was selected due to its evaluation focus on conversational agents. Traditional metrics like the SUS (Brooke et al., 1996), though valuable, may not fully capture the nuanced aspects of chatbot interactions. The CUQ, in contrast, is designed to assess these aspects, making it a more suitable tool for evaluating the overall usability and effectiveness of chatbots.

While the original CUQ questions focus on the usability and evaluation of the chatbot's interface, they barely cover linguistic aspects. Henceforth, we modified the CUQ so that it could assess the chatbot's communication style, particularly the impact of empathetic versus neutral tones, which was one of the main objectives of this study. The adapted CUQ has two sets of questions: the first 8 of them evaluate the linguistic aspects of the interactions, and the other 8 focus on the usability aspect (see Table 2).

	Question
1	The chatbot's personality was realistic and engaging.
2	The chatbot seemed too robotic.
3	The chatbot was welcoming during initial setup.
4	The chatbot seemed very unfriendly.
5	The chatbot acknowledged my feelings appropriately.
6	The chatbot ignored my concerns.
7	The chatbot used language that was considerate and supportive.
8	The chatbot communicated in a cold and distant manner.
9	I trust the information provided by the chatbot.
10	I am skeptical of the advice the chatbot gave me.
11	Chatbot responses were useful, appropriate and informative.
12	Chatbot responses were irrelevant.
13	I am satisfied with my experience interacting with the chatbot.
14	My experience interacting with the chatbot was frustrating.
15	I would recommend this chatbot to others for lifestyle change advice.
16	I would advise others against using this chatbot for lifestyle change advice.

Table 2: Adapted Chatbot Usability Questionnaire used in our user experience study.

5.3 Data Collection and Analysis

Our data, collected anonymously and remotely, consist of the questionnaire's responses and chatlogs.

To investigate the impact of the empathy condition on the CUQ scores, we conducted a one-way ANOVA with blocking, using chatbot experience, chatbot opinion, and medical chatbot opinion as block variables. Before proceeding with the analysis, the dataset underwent a rigorous process to check for normality and homogeneity of variances. Normality tests, such as Q-Q plots and histograms, were conducted to visually inspect that the CUQ scores within each group (empathetic or neutral chatbot) followed a normal distribution. Additionally, the Kolmogorov–Smirnov test was applied to confirm that the distribution of the CUQ scores do not significantly differ from a normal distribution with equal mean and deviation.

Additionally, we conducted a qualitative analysis using the chatlogs and participants' answers to the open questions in the questionnaire.

5.4 Results and Discussion of the Quantitative Analysis

5.4.1 Overall CUQ Score

The overall CUQ score comprises the results from the complete questionnaire, without any distinction between the nature of the questions. The mean overall CUQ score for the empathetic chatbot was 66.3±17.0, and 49.4±20.3 for the neutral one. Moreover, the empathetic chatbot consistently scored higher across all the scenarios, as it can be seen on Figure 2.

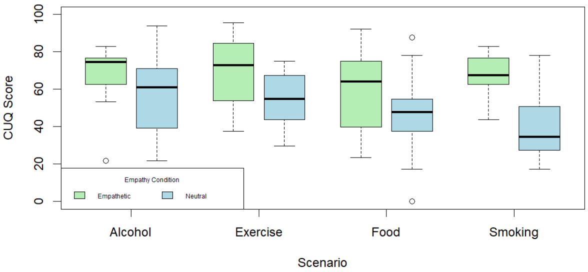


Figure 2: Overall CUQ scores per empathy condition and scenario.

The ANOVA results show that the empathy condition is highly statistically significant, with a p-value of 0.00002138 (p <0.001), whereas the block variables (chatbot opinion, medical chatbot opinion, and chatbot use frequency) are not. The ANOVA coefficients illustrate how much changing each variable modifies the CUQ score. The intercept shows us the "base case" which, in this case, it is when the condition is empathetic, the chatbot frequency of use is yearly, and the opinion towards regular and medical chatbots is uncertain. In this base case, the average CUQ score was 64.9±4.0. Then, it showcases that, if from this base case we only change the condition to neutral, without modifying all the other variables, the average CUQ score will be reduced by -17.0±3.8. This effect is significant (p <0.001). Other block variables are not statistically significant.

5.4.2 Linguistic CUQ Score

The linguistic CUQ score encompasses a subsection of scores about linguistic statements. These sentences evaluated if the chatbot's linguistic style while providing answers was perceived as welcoming, friendly and supportive by the participants. The empathetic chatbot had a mean linguistic CUQ score of 59.3±9.7, compared to 50.2±9.2 for the neutral chatbot. Similarly to the previous section, the empathetic chatbot consistently outperformed the neutral one across all scenarios, as illustrated in Figure 3.

The ANOVA reveals that the empathy condition is also highly statistically significant, with a p-value of 0.000007095 (p <0.001). Regarding the ANOVA coefficients, with the base case described in the previous section, the average CUQ score is 59.5±2.1. If the condition shifts from empathetic to neutral, without altering any other factors, the average CUQ score decreases by -9.1±2.0, a change that is statistically significant (p <0.001).

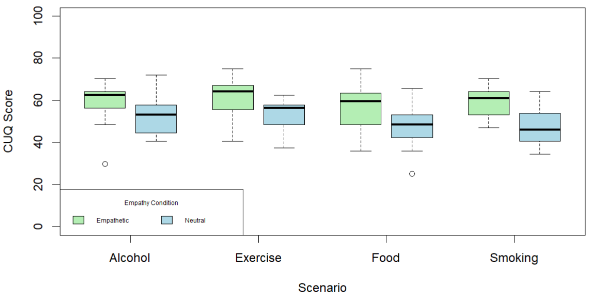


Figure 3: Linguistic CUQ scores per empathy condition and scenario.

The remaining block variables do not have a significant effect.

5.4.3 Usability CUQ Score

The usability CUQ score includes a subset of scores related to usability statements, assessing whether participants perceived the chatbot's answers as useful and relevant for lifestyle change advice. The empathetic chatbot had a mean usability CUQ score of 57.1±8.5, while the neutral chatbot scored 49.2±12.3. Consistent with previous findings, the empathetic chatbot consistently surpassed the neutral one in all scenarios, as shown in Figure 4.

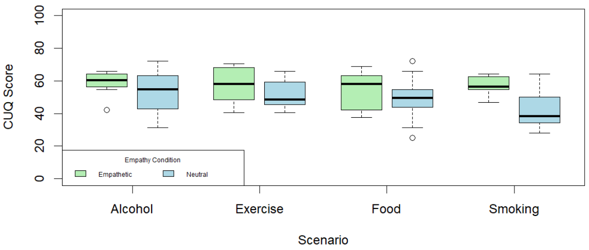


Figure 4: Usability CUQ scores per empathy condition and scenario.

The ANOVA demonstrates that the empathy condition has a highly statistically significant effect, with a p-value of 0.0003546 ($p < 0.001$). This indicates that the variation observed in the CUQ scores is unlikely to be due to chance. In the context of the base case described in the previous sections, the average CUQ score is 55.5±2.3. When the condition is shifted from empathetic to neutral, while keeping all other conditions constant, there is a notable decrease in the average CUQ score by -7.9±2.1. This decrease is statistically significant, with a p-value of less than 0.001, highlighting the impact of the empathy condition on the CUQ scores. Additionally, the analysis reveals that the remaining block variables do not have a significant effect on the CUQ scores.

5.5 Results and Discussion of the Qualitative Analysis

5.5.1 Chatbot Dialogues

The dialogue excerpts obtained from the chatlogs highlight the differences in the way neutral and empathetic chatbots respond to user queries. In the interactions with the neutral chatbot, responses were direct and factual, with no additional commentary or expression of understanding. For example, when a participant asked about fruits low in sugar, the chatbot simply listed "apples, pears, and berries" without further elaboration. This pattern is consistent across all interactions with the neutral chatbot, where the focus was on delivering concise and straightforward information.

In contrast, the empathetic chatbot provided responses that not only addressed the participants' queries but also incorporated elements of empathetic communication. The responses often began with expressions of understanding or concern, followed by advice or information that was more detailed and personalised. For instance, when a participant mentioned feeling sluggish after meals, the empathetic chatbot acknowledged the participant's feelings and provided a comprehensive answer that included suggestions for dietary adjustments and a rationale behind those suggestions.

This approach aligns with the lexical and phrasal choices associated with empathetic communication as identified by Yaden et al. (2023) and Lapointe (2014), such as the use of first and second person pronouns ("I understand that you feel..."), modal verbs ("would", "could"), and phrases that validate the user's experiences ("I hope this advice was helpful."). Furthermore, an n-gram frequency analysis of the chatlogs reveals significant differences in word usage between the empathetic and neutral chatbots. Specifically, the words identified in Yaden et al. (2023) as characteristic of empathetic communication constitute 14.04% of all unigrams produced by the empathetic chatbot, compared to 6.80% in the neutral chatbot. Similarly, the two-word phrases listed in Lapointe (2014), account for 3.27% of all bigrams generated by the empathetic chatbot, but only 0.86% in the neutral one. These differences are highly statistically significant ($p < 0.001$).

5.5.2 User Feedback

Participants' feedback further supports the contrast between the interactions with the neutral and empathetic chatbots. Users often described the responses of the neutral chatbot as "cold" and "robotic", noting the lack of empathetic engagement. One participant remarked that the chatbot's responses felt like "getting a list of Google results", which indicates that the interaction was perceived as impersonal and purely informational.

Conversely, feedback on the empathetic chatbot was generally positive, with participants appreciating the more engaging and supportive nature of its responses. Participants highlighted that the empathetic chatbot provided "useful" information and that the interaction felt "lively" and "holistic". One participant even mentioned that the chatbot's advice made them seriously consider changing their behaviour, such as reducing alcohol consumption. These comments and descriptions align with the conclusions from the previous subsection on the chatbot dialogues.

6 Conclusion

This paper aimed to investigate how the use of empathy in generated messages can affect user experience during queries about lifestyle changes.

The two primary contributions of this study are to provide insight in the impact of empathetic versus neutral tones in messages in a LLM based chatbot, and to understand user expectations in human-computer interactions - using chatbots - in the healthcare domain, especially on lifestyle changes.

The results of the first experiment show the differences between different LLMs, specifically two domain-specific and two general ones, on the different metrics fluency, coherence, groundedness and naturalness. These differences are not big, and the model with the most naturalness on the MASH-QA dataset concerning lifestyle questions - MedAlpaca - is chosen as the model to use in the second experiment.

The results of the second experiment show that empathy plays a crucial role in enhancing user satisfaction. The empathetic chatbot significantly outperformed the neutral chatbot across all dimensions measured by the Chatbot Usability Questionnaire, including overall user experience, linguistic perception, and usability ($p < 0.001$). This outcome highlights the importance of empathy in chatbot communication, especially in healthcare settings where users are likely to seek comfort and understanding.

Beyond just evaluating chatbot performance, it was essential to analyse what users expect from these interactions and how these expectations shape their experience. Results revealed that users expect healthcare chatbots to offer more than just accurate and relevant information — they expect to participate in an interaction that mirrors human conversation. The high CUQ scores for the empathetic chatbot suggest that when these expectations are met, users are more satisfied and more likely to view the chatbot as a trustworthy and effective tool for asking about health advice.

Some of the limitations of this work include that the user experiment was specifically set to 4 scenarios and the participants recruited were not searching actively for lifestyle change. Although the participants were free to use their wording, the scenarios were quite restricted. In future work, it would be nice to conduct the experiment in a more realistic setting with more participants to verify our findings. Additionally, the sample size was relatively small and homogeneous, which hinders the generalisation of the results to a broader population. For example, individuals from different educational backgrounds or age groups might prioritise straightforwardness over empathy, which could yield slightly different results over the preferred tone in messages. Future work could replicate the experiment with a larger, more diverse sample to verify whether these preferences could be applied universally or are influenced by specific demographic factors.

In summary, it is evident that the most effective healthcare chatbots are those that balance generating accurate medical information with an empathetic dialogue style. While other general linguistic capabilities are important, the success of a healthcare chatbot heavily relies on its ability to communicate empathetically and align with human conversational patterns. This project has demonstrated that incorporating empathy into chatbot design can significantly improve user experience, making these tools more appealing and effective in supporting lifestyle changes and health-related decision-making.

Acknowledgments

This project has been supported by the Department of Data Science at TNO.

References

Parham Amiri and Elena Karahanna. 2022. Chatbot use cases in the Covid-19 public health response. *Journal of the American Medical Informatics Association*, 29(5):1000–1010.

Wahied Khawar Balwan and Sachdeep Kour. 2021. Lifestyle Diseases: The Link between Modern Lifestyle and Threat to Public Health. *Saudi Journal of Medical and Pharmaceutical Sciences*, 7(4):179–184.

Maaike H. T. de Boer, Jasper van der Waa, Sophie van Gent, Quirine T. S. Smit, Wouter Korteling, Robin M. van Stokkum, and Mark Neerincx. 2023. A contextual Hybrid Intelligent System Design for Diabetes Lifestyle Management. In *Proceedings of the Fourteenth International Workshop on Modelling and Representing Context (MRC 2023).*

John Brooke et al. 1996. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Jacky Casas, Timo Spring, Karl Daher, Elena Mugellini, Omar Abou Khaled, and Philippe Cudré-Mauroux. 2021. Enhancing conversational agents with empathic abilities. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 41–47.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *arXiv preprint arXiv:2311.16079.*

Benjamin M.P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. 2016. Empathy: A Review of the Concept. *Emotion Review*, 8(2):144–153.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv preprint arXiv:2304.08247.*

Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. https://doi.org/10.1145/3335082.3335094 Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, page 207–214.

Marco Iacoboni. 2005. Understanding Others: Imitation, Language, and Empathy. In Susan Hurley and Nick Chater, editors, *Perspectives on Imitation: Vol. 1. Mechanisms of Imitation and Imitation in Animals*, pages 77–99. The MIT Press, Cambridge, MA.

Stephanie Lapointe. 2014. A Corpus Study of the Verbal Communication of Empathy/Sympathy by Anglophone Nurses in Quebec. Master's thesis, University of Quebec at Chicoutimi, UQAC Repository.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Bingjie Liu and S Shyam Sundar. 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, 21(10):625–636.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634.*

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. *AI at Meta Blog.*

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv:2303.13375.*

OpenAI. 2024. GPT-4 Technical Report. *arXiv:2303.08774.*

Renate L. E. P. Reniers, Rhiannon Corcoran, Richard Drake, Nick M. Shryane, and Birgit A. Völlm. 2011. The QCAE: A Questionnaire of Cognitive and Affective Empathy. *Journal of Personality Assessment*, 93(1):84–95.

Yi Shan, Meng Ji, Wenxiu Xie, Xiaobo Qian, Rongying Li, Xiaomin Zhang, and Tianyong Hao. 2022. Language Use in Conversational Agent–Based Health Communication: Systematic Review. *Journal of Medical Internet Research*, 24(7):e37403.

David B. Yaden, Salvatore Giorgi, Matthew Jordan, Anneke Buffone, Johannes C. Eichstaedt, H. Andrew Schwartz, Lyle Ungar, and Paul Bloom. 2023. Characterizing Empathy and Compassion Using Computational Linguistic Analysis. *Emotion*, 24(1):106–115.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. *arXiv preprint arXiv:2210.07197*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. *arXiv:2311.05112*.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question Answering with Long Multiple-Span Answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849.