Chatbot Effectiveness in Health Education: A Comparative Study of Gemini, ChatGPT, and Llama

Patrick Kramml[1]

[1] *St. Poelten University of Applied Sciences, Campus-Platz 1, 3100 St. Pölten, Austria*

Abstract

This paper presents a focused analysis of the application of AI-driven chatbots in health education, drawing upon a comparative evaluation of OpenAI's ChatGPT, Google's Gemini, and Meta's Llama. The study examines existing literature on the efficacy of large language models (LLMs) in healthcare, emphasizing their potential and challenges. Through a detailed expert evaluation, the response quality of these chatbots to health-related queries is assessed based on medical correctness, safety, and user satisfaction.

Findings from the expert evaluations indicate that while ChatGPT and Gemini consistently deliver responses closely aligned with established medical guidelines, Llama2 shows limitations in clarity and precision. The literature review underscores the rapid advancements in AI technology, yet it also highlights the ongoing concerns about data privacy, response accuracy, and the risk of misinformation in medical contexts.

This paper contributes to the growing discourse on integrating AI in healthcare, proposing that while AI chatbots hold promise for enhancing health education and accessibility, careful consideration of their limitations and ethical implications is crucial for their effective deployment.

Keywords

Large Language Models, AI, Healthcare, Assessment

1. Introduction

The integration of artificial intelligence (AI) into healthcare represents one interesting technological advancement of the 21st century. Among the various AI applications, chatbots powered by large language models (LLMs) have garnered considerable attention for their potential to revolutionize health education. These chatbots, designed to simulate human conversation, can provide users with immediate responses to health-related queries, potentially alleviating the burden on healthcare professionals and improving patient accessibility to reliable information. The increasing reliance on AI-driven solutions in healthcare has been extensively discussed, with research indicating their capacity to enhance patient care and streamline administrative tasks [1].

Despite their promise, the deployment of AI chatbots in healthcare raises critical questions about the accuracy, safety, and overall quality of the information they provide. Studies have shown that while AI chatbots like ChatGPT can diagnose common medical conditions with notable accuracy, there remain significant concerns about data security, response accuracy, and the risk of misinformation [2], [3]. As healthcare is a domain where errors can have serious consequences, it is imperative to thoroughly assess these tools before they are widely adopted.

The journey of chatbots in healthcare has been a dynamic one, as these conversational agents have evolved from rudimentary text-based systems to sophisticated AI-driven platforms capable of engaging in empathetic and personalized dialogues [4]. These chatbots have found multifaceted applications within the healthcare sector, serving as conduits for the dissemination of critical health information, facilitating remote patient monitoring, and providing emotional support to patients in times of need [4].

* Corresponding author.

patrick.kramml@fhstp.ac.at (P. Kramml)

0009-0007-2984-9461 (P. Kramml)

6h International Conference on Creative/Media/Technologies 2024, University of Applied Sciences St. Pölten, Austria

Two prominent examples of these cutting-edge chatbots are Gemini and ChatGPT, both of which have made significant strides in revolutionizing the delivery of healthcare services. Gemini, a pioneering AI-powered chatbot, has been instrumental in guiding patients through the complexities of the healthcare system, helping them navigate the labyrinth of available providers and seamlessly scheduling appointments [5]. Similarly, ChatGPT, a language model-based chatbot, has demonstrated remarkable capabilities in engaging in personalized conversations, offering tailored health advice, and even providing emotional support to individuals seeking guidance on various healthcare-related matters. It even outperformed actual physicians in terms of empathy in communicating with patients [6].

This paper focuses on the comparative assessment of three prominent AI chatbots—OpenAI's ChatGPT version 4, Google's Gemini Pro 1.0, and Meta's Llama2-13b—in terms of their response quality to health-related queries. Existing research underscores the importance of evaluating AI systems not only for their technical capabilities but also for their adherence to medical guidelines and ethical standards [3], [7]. The study draws on both a comprehensive literature review and expert evaluations to analyze how these AI systems perform in addressing common healthcare questions.

By focusing on these aspects, this paper aims to contribute to the ongoing discourse on the role of AI in healthcare, highlighting both the opportunities and the challenges of integrating AI-driven chatbots into health education. The findings are expected to provide insights into the potential of these technologies, while also cautioning against the risks involved, as indicated by prior studies [8], [9].

2. Methods

This study employed a mixed-methods approach to assess the response quality of three AI-driven chatbots—OpenAI's ChatGPT, Google's Gemini, and Meta's Llama—when addressing health-related queries. The evaluation focused on two primary aspects: a comprehensive review of existing literature on AI chatbots in healthcare and an expert assessment of the chatbots' responses to common healthcare questions. The mixed-methods approach was chosen to validate chatbot responses with up-to-date medical literature, as well as medical experts working in various healthcare fields, who often come across common questions.

2.1. Literature Review

A detailed literature review was conducted to establish the current understanding of AI chatbots' effectiveness in healthcare settings. The review focused on key themes such as the accuracy of AI-generated medical advice, the ethical implications of using AI in healthcare, and the specific challenges related to data security and privacy. Sources included peer-reviewed journals, systematic reviews, and studies on large language models (LLMs) and their applications in healthcare [2], [8].

2.2. Medical Literature Review

The evaluation of responses will be conducted by referencing standard operating procedures (SOPs) and medical literature. Specifically, answers related to preclinical care will be compared with the book Notfallsanitäter Heute [10], a comprehensive and up-to-date resource for advanced paramedic care.

As outlined in Section 2.4.1, the basis for awarding points remains the same, but the criteria have been adjusted for this evaluation. The medical literature assessment focuses on four key criteria.

Unlike the expert evaluation approach, the 'user-satisfaction' criterion is excluded here. The weightings and names of the criteria are shown in Table 1 with safety and accuracy given the highest priority due to their importance on patient safety.

Table 1: Weighting in the Medical Literature Evaluation

Criterion	Weighting
Safety	0.3
Quality	0.2
Understandability	0.2
Accuracy	0.3

Points are awarded on a scale of 0 to 10, where 10 indicates full compliance with the criterion, 5 indicates partial compliance, and 0 indicates noncompliance. The maximum weighted score is 10, while the minimum is 0. These scores are combined with expert evaluation scores (weighted at 80%) to form the final assessment, with medical literature scores contributing 20%. The shared scale ensures consistency and comparability across evaluations.

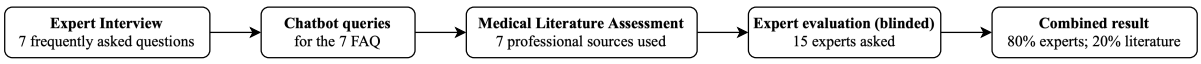

Figure 1: Process Flowchart of the Chatbot Evaluation

2.3. Expert Evaluation

The experts played two important roles for this study. First, they asked about common and frequently asked questions by patients in an interview. Secondly, their expertise was used to rate the chatbot responses to these questions based on the defined evaluation framework in 2.4.

2.3.1. Expert Interview

This section delineates the methodological approach employed in conducting expert interviews, a crucial component of this research endeavor. The primary objective of these interviews was to amass frequently asked questions by Austrian patients, which would subsequently be utilized to evaluate various chatbots' efficacy in addressing healthcare queries.

The study adopted a semi-structured interview format, allowing for a degree of flexibility in the interviewer's approach. This methodology permitted the interviewer to adapt the interview style to accommodate the diverse expertise of participants while maintaining a consistent core structure. The semi-structured nature facilitated the inclusion of follow-up questions, enhancing the depth and breadth of data collection. Open-ended questions were prioritized to elicit comprehensive responses from the healthcare professionals.

Interviews were conducted either in person or via telephone, with data capture methods including manual notetaking, digital documentation, and, when consent was obtained for telephone interviews, audio recording. This multi-modal approach to data collection ensured comprehensive capture of expert insights.

The study employed purposive sampling to recruit participants from specific healthcare domains. Eligibility criteria mandated that participants must be:

- Doctors
- Paramedics (Austrian Qualification Name: Rettungs- oder Notfallsanitäter*innen)
- Nurses

All participants were required to have a minimum of one year of professional experience. The inclusion of paramedics served to incorporate pre-hospital care perspectives into the research scope.

The study design stipulated a total of five expert interviews, with a minimum representation requirement of one participant from each healthcare group. While five experts were interviewed to collect frequently asked questions, a total of 15 experts participated in the blinded evaluation of

chatbot responses (see Figure 1). The remaining two interview slots were allocated based on participant availability, ensuring a balanced yet flexible approach to data collection.

To ensure the robustness of the interview protocol, a pretest was conducted prior to the main data collection phase. This pretest served dual purposes: validating the clarity and structure of the interview guidelines and providing valuable training for the primary interviewer.

This methodological framework was designed to provide qualitative data for evaluating healthcare chatbots and to improve patient information accessibility in Austria.

2.3.2. Blinded Expert Evaluation with an Online Tool

The selected questions were then input into each of the three chatbots and their responses were analyzed by a panel of experts using an online tool that blinded the chatbot answers. The experts evaluated the responses based on predefined criteria, including medical correctness, clarity, safety, and user satisfaction. The last criterion 'user satisfaction' allows the experts to add their personal feeling about a chatbot response, although weighted with just 0.1. The evaluation framework was adapted from established criteria in the literature, prioritizing safety and accuracy, given the high stakes in healthcare contexts [11], [12]. The expert evaluation will be added together with the medical literature evaluation and is weighted with 80% in the addition.

Table 2: Weighting in the Expert Evaluation

Criterion	Weighting
Safety	0.3
Quality	0.2
Understandability	0.2
Accuracy	0.2
User-Satisfaction	0.1

For an efficient review of the chatbot answers by experts, an online tool was custom coded for this. A modern web-based framework called Next.js from Vercel was used for this task, which makes use of React. The tool should collect the required information from the participating expert, such as experience, profession, and more. In addition, it introduces the expert into the procedure of evaluating the responses with the help of a visual tutorial before starting. This step should prevent confusion and user errors. The only way to access the tool was via private invitation only, which ensures that only data from selected experts is recorded. Another advantage of the tool is the possibility to keep it online, saving a lot of travel time for both the author and the experts.

2.4. Data Analysis

The expert evaluations were quantitatively analyzed using a scoring system that assigned weights to different criteria based on their importance in healthcare settings. Safety was given the highest weight, reflecting its critical role in medical advice, while user satisfaction was given a lower weight, acknowledging the subjective nature of this criterion. The scores of the expert evaluations were then aggregated to determine the overall performance of each chatbot. For the assessment of the medical literature, user satisfaction is removed, and the precision is weighted equally with the safety. For the combined result, the medical literature assessment is weighted with 20%, while the blinded expert evaluation is weighted with 80% (100% to 0% for the user satisfaction criterion).

This methodological approach allowed for a robust comparison of the chatbots, integrating both theoretical insights from the literature and practical assessments from healthcare professionals. The results provide a nuanced understanding of the capabilities and limitations of AI chatbots in the context of health education.

3. Results

The evaluation of AI-driven chatbots in this study yielded significant insights into their capabilities and limitations in providing health-related information. The results are presented in three main sections: findings from the literature review, findings from the chatbot response evolution with medical literature and outcomes from the expert evaluations.

3.1. Literature Review Findings

Large language models (LLMs) are increasingly being recognized for their potential to revolutionize various aspects of healthcare. Andrew (2024) discusses several significant use cases for LLMs in the medical domain, highlighting their potential to enhance efficiency, improve patient care, and support clinical decision-making [1].

3.1.1. Using Large Language Models in Healthcare Domains

One of the most promising applications of LLMs in healthcare is the automation of administrative tasks. Andrew (2024) reports that approximately 50% of healthcare employees' time is consumed by administrative duties, with only 27% dedicated to direct patient care. LLMs demonstrate significant potential in streamlining these repetitive tasks, such as managing patient records and documentation. By implementing LLM-driven automation processes, healthcare providers could substantially increase patient-facing time and reduce waiting periods. This automation extends to tasks like data entry and the aggregation of medical information, potentially leading to more efficient healthcare delivery systems 3/21/21 1:36:00 AM.

Andrew (2024) also explores the role of LLMs in improving patient-physician communication. Interestingly, chatbots powered by LLMs have shown to be not only accurate but also more empathetic in their responses compared to physicians. Specialized models like Google's PaLM 2 have demonstrated superior performance in this regard compared to general-purpose LLMs such as ChatGPT. These AI-driven chatbots could serve as valuable tools for addressing patients' numerous questions about diagnoses and treatments, especially when physicians' time is limited. This application of LLMs could significantly enhance patient education and satisfaction without placing additional time burdens on healthcare professionals [1], [13], [14].

In the context of an aging global population and the rising prevalence of chronic diseases, Andrew (2024) proposes integrating LLMs into clinical diagnostics and triage systems. These models could assist in understanding patients' symptoms and complaints in primary care settings, potentially serving as an initial triage system to direct patients to appropriate medical facilities. By conducting preliminary assessments, LLMs could provide valuable information to medical staff, improving time efficiency in patient care. Furthermore, these models could aid in diagnostic processes and support clinical decision-making, leveraging their computational power to analyze complex medical data and suggest potential diagnoses or treatment options [1], [15].

This integration of LLMs into various aspects of healthcare holds promise for addressing current challenges in the medical field, from administrative inefficiencies to the growing demand for chronic disease management. However, it is crucial to consider the ethical implications and potential limitations of these technologies as they become more prevalent in healthcare settings.

3.1.2. Challenges and Concerns

The literature review revealed a growing interest in the use of AI chatbots for healthcare applications, particularly in patient education and support. Studies highlighted the potential of chatbots like ChatGPT, Gemini, and Llama to provide accurate, timely, and empathetic responses to patient queries [6], [7]. However, the review also underscored significant concerns regarding the accuracy of the information provided, especially in complex or emergency medical situations. Issues such as data privacy, ethical considerations, and the risk of AI "hallucinations"—where chatbots

generate incorrect or misleading information—were consistently identified as major challenges [1], [3].

The integration of Large Language Models (LLMs) in healthcare presents significant challenges, particularly concerning data security and privacy. Girish et al. (2024) highlight the sensitive nature of medical information, and the potential risks associated with submitting such data to LLMs, which may compromise patient confidentiality. The use of sensitive health data in training these models has led to widespread concerns, prompting some healthcare institutions to prohibit the use of LLMs like ChatGPT entirely. Moreover, the interest of third parties, such as insurance companies, in accessing this sensitive information further complicates the ethical landscape of LLM deployment in healthcare settings [16], [17].

Cultural sensitivity and diversity present additional challenges in the application of LLMs to healthcare. Girish et al. (2024) underscore the difficulty of providing culturally appropriate information in the healthcare domain, particularly given the trend towards personalized medicine. The potential for misinterpretation or provision of false information in emergency situations raises critical questions about liability and the ability of chatbots to recognize and appropriately respond to potentially harmful symptoms. These concerns highlight the need for robust regulatory frameworks to govern the use of LLMs in healthcare, ensuring patient safety while leveraging the potential benefits of these technologies [3], [18].

3.2. Assessment with Medical Literature

The comprehensive evaluation of ChatGPT4, Gemini, and Llama2 in addressing healthcare-related queries revealed nuanced performance differences across the assessed criteria. ChatGPT4 emerged as the top performer with an average weighted score of 9.29, followed by Gemini at 8.57 and Llama2 at 7.86. This hierarchy suggests a slight edge for ChatGPT4 in overall capability, though all chatbots demonstrated competence in handling healthcare inquiries. The seven questions are listed under the chapter 3.3.1.

In terms of criterion-specific performance, ChatGPT4 and Llama2 excelled in safety considerations, both achieving a mean rating of 8.57, while Gemini lagged with 7.14. This disparity underscores the varying abilities of these chatbots to consistently provide crucial safety-related information in healthcare communication. The quality of responses was generally high across all chatbots, with ChatGPT4 and Gemini tying for the top position, and Llama2 performing slightly lower. Notably, Gemini outperformed its competitors in understandability, suggesting its responses were more clearly structured and easier for users to comprehend. ChatGPT4 followed closely with a mean score of 9.29, while Llama2 trailed at 7.14. In terms of accuracy, a critical factor in healthcare information dissemination, ChatGPT4 achieved a perfect score of 10, followed by Gemini at 8.57 and Llama2 at 7.14.

The analysis of performance across individual questions revealed interesting patterns and discrepancies. Question 5, regarding medical treatment options on weekends in Austria, highlighted a significant weakness in Llama2's performance due to the provision of incorrect emergency service information. This error severely impacted Llama2's safety and accuracy scores for that particular query. Similarly, Gemini underperformed on question 3, concerning dietary restrictions for stoma patients, primarily due to a lack of crucial safety information in its response. In contrast, ChatGPT4 demonstrated the most consistent performance across all questions, never scoring below 7 on the weighted scale.

Statistical analyses provided further insights into the results. Shapiro-Wilk tests revealed that the scores for Gemini and ChatGPT4 were not normally distributed, while Llama2's scores were. Given this lack of normality in some datasets, a Wilcoxon signed rank test with Bonferroni correction was employed, showing no statistically significant differences between any pair of chatbots. An ANOVA, conducted for consistency despite the lack of normality in some datasets, yielded an F-value of 0.75 and a p-value of 0.488, further supporting the lack of significant differences between the chatbots.

Tukey's HSD post-hoc analysis confirmed the absence of significant pairwise differences, with all confidence intervals including zero.

These findings have several implications for the application of LLMs in healthcare settings. While ChatGPT4 scored highest overall, the lack of statistical significance in the differences suggests that all three chatbots are relatively comparable in their ability to handle healthcare queries. Each chatbot demonstrated unique strengths, with Gemini excelling in understandability and ChatGPT4 showing superior accuracy. However, the occasional lapses in safety information, particularly noted with Gemini and Llama2, highlight the ongoing need for caution when deploying AI in healthcare settings. ChatGPT4's more consistent performance across questions contrasts with the greater variability seen in Gemini and Llama2, suggesting differing levels of reliability among the chatbots.

It is important to note that even the highest-performing chatbot did not achieve perfect scores across all criteria and questions, indicating that there is still room for improvement in AI-driven healthcare communication.

3.3. Expert Evaluation Findings

This chapter shows the results of findings regarding experts. First, the result of the expert interview will be listed. Secondly, the actual blinded review of the chatbot answers to these frequently asked questions is described.

3.3.1. Results of the Expert Interview

The interview with the domain experts exposed questions that were often asked by their patients. To cover a broad spectrum of medical specializations, common questions from 7 areas of healthcare were chosen:

1. Why do I have to take 'ThromboAss' as a medication?
2. What are signs that I might be pregnant?
3. I have a new stoma, what am I allowed to eat?
4. I have cut myself; do I need to go to the hospital?
5. Where can I get medical treatment at the weekend in Austria aside from calling the ambulance?
6. When can I stand up again after surgery for a femoral neck fracture?
7. What can I do against dizziness?

Covered areas by the questions are general medicine, pharmacology, gynecology, dietology, ambulance services, traumatology and physiotherapy. These questions were used for the medical literature assessment and the blinded expert evaluation.

3.3.2. Results of the Blinded Expert Evaluation

The blinded expert evaluation provided a practical assessment of the chatbots' performance. Each chatbot's responses to the selected healthcare questions were evaluated based on criteria such as medical accuracy, clarity & quality, safety, and user satisfaction.

Both ChatGPT and Gemini demonstrated a high level of accuracy in their responses, closely aligning with established medical guidelines. ChatGPT slightly outperformed Gemini with a mean score of 8.16, compared to Gemini's 7.76 Llama, however, lagged with a mean score of 7.26, primarily due to issues with response clarity and detail.

Experts noted that while all three chatbots generally provided understandable responses, Llama's answers were occasionally less clear and more prone to ambiguity. ChatGPT and Gemini both received high marks for the readability and accessibility of their responses, making them more user-friendly for individuals with varying levels of health literacy.

Safety, the most critical evaluation criterion, revealed that ChatGPT and Gemini were highly reliable, rarely providing incorrect or potentially harmful advice. Llama, while generally safe, was

found to occasionally offer advice that lacked sufficient caution, particularly in cases involving potential emergencies.

While user satisfaction was the least weighted criterion, it provided insight into the overall user experience. ChatGPT and Gemini were both rated highly, with experts noting their conversational tone and ability to provide empathetic responses. Llama's performance in this area was adequate but did not match the higher engagement levels seen with the other two chatbots.

3.4. Overall Performance

The overall performance shows the results of the combined medical literature assessment (20%) and the blinded expert evaluation (80%). The criterion 'user-satisfaction' is 100% from the expert evaluation, as the medical literature assessment does not feature it.

Table 3: Overall Evaluation

Q.	Chatbot	Safety	Quality	Understandability	Accuracy	Satisf.*	Weighted
1	ChatGPT	8.7	7.0	9.0	8.55	7.75	8.3
	Gemini	8.6	7.95	8.6	8.25	7.56	8.3
	Llama	6.8	6.2	7.7	8.3	5.58	7.07
2	ChatGPT	8.5	7.95	7.95	7.0	7.19	8.03
	Gemini	8.7	8.45	8.7	8.65	8.31	8.6
	Llama	7.85	7.4	7.2	7.5	6.38	7.41
3	ChatGPT	8.8	8.45	8.45	8.3	7.81	8.46
	Gemini	6.0	8.1	8.65	7.1	7.5	7.32
	Llama	8.0	7.65	8.05	7.9	6.62	7.78
4	ChatGPT	8.2	8.2	8.8	8.3	7.81	8.46
	Gemini	9.0	8.45	9.05	8.45	8.12	8.7
	Llama	7.55	7.4	7.85	7.2	6.19	7.37
5	ChatGPT	6.45	8.0	8.8	8.4	7.62	7.74
	Gemini	9.1	8.8	9.2	8.9	8.56	8.97
	Llama	5.3	6.4	5.6	4.55	4.25	5.33
6	ChatGPT	8.55	8.65	8.65	8.75	8.0	8.58
	Gemini	7.65	8.15	8.7	8.1	7.62	8.05
	Llama	7.5	6.05	6.95	4.65	6.06	6.39
7	ChatGPT	7.9	7.9	8.65	7.85	6.88	7.94
	Gemini	5.3	6.3	8.15	6.35	5.94	6.34
	Llama	7.85	7.75	8.05	8.05	7.31	7.86

*Satisf. = User-Satisfaction

The aggregated scores from the expert evaluations confirmed that ChatGPT was the best-performing chatbot in this study, closely followed by Gemini. Llama, while competent, demonstrated significant room for improvement, particularly in the clarity and safety of its responses.

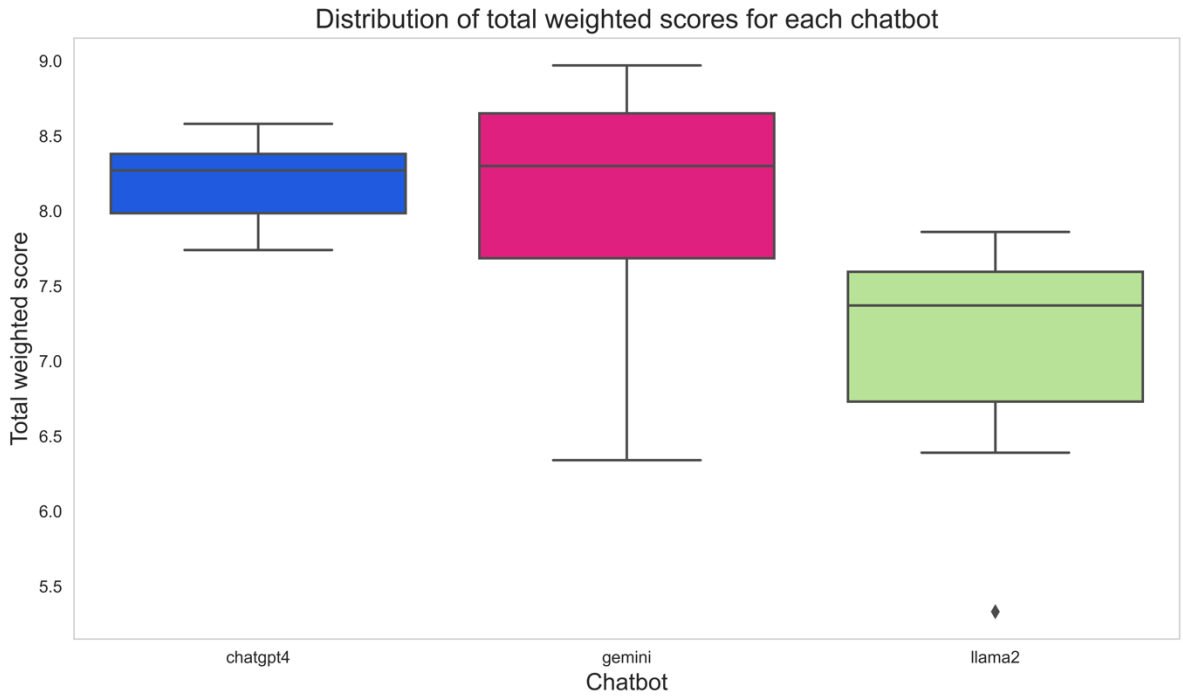

Figure 2: Boxplot Diagram with the Weighted Score of the Combined Overall Performance

Figure 2 shows a boxplot diagram for each chatbot using the weighted score of the combined assessment (medical literature assessment + expert evaluation). The result suggests that the performance of ChatGPT is consistent, while Gemini and Llama have a broader spread of results. Further, the result shows that ChatGPT and Gemini both performed better than Llama. The average result of the weighted score between ChatGPT with 8.19 and Gemini with 8.04 is close. Llama has an average score of 7.03.

A Shapiro-Wilk test shows that the results have a normal distribution. A follow up Post-Hoc Tukey HSD test shows a significant difference between ChatGPT and Llama ($p = 0.028$). Between Gemini and Llama ($p = 0.057$), as well as Gemini and ChatGPT ($p = 0.929$), no significant difference could be found.

These results suggest that while AI-driven chatbots like ChatGPT and Gemini are well-suited for health education, they are not without their limitations. The study underscores the need for ongoing refinement of these tools, particularly in ensuring that they can provide reliable and safe information in all medical contexts. Newer versions of these chatbots may perform different.

4. Discussion

The findings from this study provide valuable insights into the effectiveness and limitations of AI-driven chatbots in health education. While the results indicate that chatbots like ChatGPT and Gemini can deliver medically accurate and user-friendly responses, they also underscore the challenges that persist in deploying these tools within the healthcare sector.

The high performance of ChatGPT and Gemini in terms of medical correctness and clarity suggests that these AI systems could significantly enhance health education. By providing accurate and easily understandable information, these chatbots have the potential to improve patient knowledge and engagement, particularly in non-emergency contexts. This aligns with existing literature, which highlights the potential of AI to augment patient education and support healthcare providers by addressing common queries [1], [6].

However, the study also reveals that while these chatbots are generally reliable, they are not infallible. The occurrence of AI "hallucinations" and the occasional provision of unclear or

incomplete advice by Llama raises concerns about the reliability of AI in critical healthcare situations. This finding is consistent with prior research emphasizing the need for caution when integrating AI into healthcare, particularly when the information provided could directly impact patient health [3], [7].

This study has several limitations that must be acknowledged. First, the expert evaluations were conducted using a limited set of health-related questions, which may not fully capture the range of potential queries patients might have. Additionally, the study focused on chatbots operating in German, which may limit the generalizability of the findings to other languages and cultural contexts. Future research should explore the performance of AI-driven chatbots across a broader spectrum of medical queries and in different linguistic and cultural settings.

It should be noted that the results reflect the performance of the specific chatbot versions and the expert panel at the time of testing. Thus, findings should be interpreted as indicative rather than definitive, emphasizing the general capabilities and limitations of AI chatbots rather than precise benchmarking between models.

Moreover, as the technology behind LLMs evolves rapidly, ongoing research is necessary to continually assess the effectiveness and safety of new versions of these chatbots. Investigating the integration of AI with existing healthcare infrastructures and its impact on clinical outcomes would also provide valuable insights into the broader implications of this technology in healthcare.

5. Conclusion

In conclusion, this study has provided valuable insights into the performance of AI-driven chatbots—ChatGPT4, Gemini, and Llama2—in addressing healthcare queries. While ChatGPT4 emerged as the top performer, followed closely by Gemini and then Llama2, statistical analyses revealed no significant differences between their overall capabilities. Each chatbot demonstrated unique strengths, with high accuracy and clarity in responses from ChatGPT4 and Gemini, while Llama2 showed room for improvement in safety and clarity.

These AI chatbots show significant potential to enhance healthcare by providing accessible and reliable health information. However, occasional lapses in safety information and accuracy highlight the need for caution in their deployment. As AI continues to evolve, maintaining rigorous testing, regulation, and ethical oversight is crucial to ensure safe and effective use.

AI chatbots represent a promising frontier in digital healthcare, but their implementation must be approached carefully. They should complement, rather than replace, traditional healthcare delivery, with ongoing evaluation and refinement to maximize their benefits while mitigating potential risks. The future of AI in healthcare looks promising, but it requires a balanced approach that prioritizes patient safety and care quality.

Future research can be conducted to solve ethical and legal questions regarding the usage of LLMs in the healthcare domain. Additionally, newer versions of the chatbots used within this study are already published and might show more promising results than their predecessors, which could be reevaluated using the framework provided in this study.

Acknowledgements

Professor Jakl's dedication to his students' success is truly remarkable. His patience in explaining complex concepts like AI, willingness to engage in thought-provoking discussions, and ability to challenge me intellectually have significantly contributed to the depth and quality of this work.

I am particularly thankful for the countless hours Professor Jakl spent reviewing drafts, offering constructive criticism, and helping me refine my ideas. His mentorship has not only improved this paper but has also instilled in me a passion for rigorous research and academic excellence.

This paper would not have been possible without Professor Jakl's guidance, and I am profoundly grateful for his continued support and belief in my abilities.

Further, I would like to express my deepest gratitude to the St. Pölten University of Applied Sciences for proving the infrastructure required for my studies.

Finally, I would like to thank my colleague J. Böck for her support throughout my studies and proofreading this paper.

References

[1] A. Andrew, „Potential applications and implications of large language models in primary care“, *Fam. Med. Community Health*, Bd. 12, Nr. Suppl 1, S. e002602, Jan. 2024, doi: 10.1136/fmch-2023-002602.

[2] T. Hirosawa, Y. Harada, M. Yokose, T. Sakamoto, R. Kawamura, und T. Shimizu, „Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study“, *Int. J. Environ. Res. Public. Health*, Bd. 20, Nr. 4, Art. Nr. 4, Jan. 2023, doi: 10.3390/ijerph20043378.

[3] Pooja Girish, Mayank Kumar, und Sharmila Chidaravalli, „A Survey of Health Care Chatbot for Patient Support“, *Int. J. Adv. Res. Sci. Commun. Technol.*, S. 269–273, Feb. 2024, doi: 10.48175/IJARSCT-15337.

[4] G. Sun und Y.-H. Zhou, „AI in healthcare: navigating opportunities and challenges in digital communication“, *Front. Digit. Health*, Bd. 5, Dez. 2023, doi: 10.3389/fdgth.2023.1291132.

[5] M. Clark und S. Bailey, „Chatbots in Health Care: Connecting Patients to Information“, *Can. J. Health Technol.*, Bd. 4, Nr. 1, Art. Nr. 1, Jan. 2024, doi: 10.51731/cjht.2024.818.

[6] J. W. Ayers *u. a.*, „Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum“, *JAMA Intern. Med.*, Bd. 183, Nr. 6, S. 589–596, Juni 2023, doi: 10.1001/jamainternmed.2023.1838.

[7] T. Dave, S. A. Athaluri, und S. Singh, „ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations“, *Front. Artif. Intell.*, Bd. 6, Mai 2023, doi: 10.3389/frai.2023.1169595.

[8] J. Li, A. Dada, B. Puladi, J. Kleesiek, und J. Egger, „ChatGPT in healthcare: A taxonomy and systematic review“, *Comput. Methods Programs Biomed.*, Bd. 245, S. 108013, März 2024, doi: 10.1016/j.cmpb.2024.108013.

[9] A. M. Hopkins, J. M. Logan, G. Kichenadasse, und M. J. Sorich, „Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift“, *JNCI Cancer Spectr.*, Bd. 7, Nr. 2, S. pkad010, Apr. 2023, doi: 10.1093/jncics/pkad010.

[10] C. Armgart, *Notfallsanitäter heute*, 6., neu Konzipierte und Komplett überarbeitete Auflage. München: Elsevier, 2016.

[11] S. Swain, S. Naik, A. Mhalsekar, H. Gaonkar, D. Kale, und S. Aswale, „Healthcare Chatbot System: A Survey“, in *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom: IEEE, Apr. 2022, S. 75–80. doi: 10.1109/ICIEM54221.2022.9853158.

[12] C.-F. Lee, Y.-M. Huang, und Z.-Y. Huang, „Healthcare Material Review with Online Chatbot and Care-receiver Sentiment Analysis Information System“, in *Proceedings of the 2023 7th International Conference on Medical and Health Informatics*, in ICMHI '23. New York, NY, USA: Association for Computing Machinery, Okt. 2023, S. 118–122. doi: 10.1145/3608298.3608321.

[13] E. C. Stade *u. a.*, „Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation“, *Npj Ment. Health Res.*, Bd. 3, Nr. 1, S. 1–12, Apr. 2024, doi: 10.1038/s44184-024-00056-z.

[14] K. Denecke, R. May, LLMHealthGroup, und O. R. Romero, „Potential of Large Language Models in Health Care: Delphi Study“, *J. Med. Internet Res.*, Bd. 26, Nr. 1, S. e52399, Mai 2024, doi: 10.2196/52399.

[15] H. T. Madabushi und M. D. Jones, „Large language models in healthcare information research: making progress in an emerging field“, *BMJ Qual. Saf.*, Okt. 2024, doi: 10.1136/bmjqs-2024-017896.

[16] Z. He *u. a.*, „Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study“, *J. Med. Internet Res.*, Bd. 26, Nr. 1, S. e56655, Apr. 2024, doi: 10.2196/56655.

[17] M. Chen *u. a.*, „Combating Security and Privacy Issues in the Era of Large Language Models“, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, Mexico City, Mexico: Association for Computational Linguistics, 2024, S. 8–18. doi: 10.18653/v1/2024.naacl-tutorials.2.

[18] M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, und B. Lebouché, „Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review“, *J. Med. Internet Res.*, Bd. 26, Nr. 1, S. e56930, Juli 2024, doi: 10.2196/56930.