

Comparative Evaluation of Generative AI Chatbots Through User-Centered Survey Analytics and Multi-Dimensional Performance Assessment Techniques

Hlaing Htake Khaung Tin¹, Khin Myo Myo Minn², Aye Chan Myae³, Thin Yanant⁴, April Thet Su⁵

Faculty of Information Science

University of Information Technology Yangon, Myanmar

¹hlainghtakekhaungtin@gmail.com, ⁵aprilthetsu@gmail.com

Abstract: *In this study, comparative analysis between state-of-the-art generative AI chatbots is done by employing survey-based analytics and performance measurements. This research focuses on the comparative analysis of generative AI chatbots through measuring response accuracy, contextual awareness, interaction capabilities, usability, response time, and user satisfaction levels. An extensive survey study was conducted involving 30 participants who have previous experience using AI chatbot services for educational, business, and general communication purposes. Data collection was performed using quantitative measures including survey questionnaires, and then data analysis was conducted through descriptive statistics and comparison of performance indicators. From the survey results, current AI chatbots offer more conversational intelligence and context-awareness than rule-based conversational systems. Among the examined chatbots, some chatbots managed to obtain a greater degree of user satisfaction because of their capabilities of producing coherent, contextually appropriate, and accurate responses. On the other hand, issues such as those of misinformation, inconsistency in complex reasoning, and contextual constraints were also discovered. It shows how significant it is to adopt an evaluation strategy that is based on users when examining conversational AI technology. This paper provides useful insights about the advantages and disadvantages of existing chatbot technologies.*

Keywords: Generative AI, Chatbots, User-Centered Evaluation, Survey Analysis, Comparative Performance Assessment.

I. Introduction

The advent of AI technology has greatly impacted on the relationship between humans and digital platforms. One of the recent innovations that has revolutionized digital communications is generative AI chatbots with large language models (LLMs). Intelligent dialogue agents capable of interpreting natural language, providing human-like replies, and participating in complex interactions in numerous fields, including education, health care, business, customer services, software development, and research, are among the many applications of AI chatbots. The latest examples of AI chatbots include OpenAI ChatGPT, Google Gemini, Anthropic Claude, and Microsoft Copilot. These sophisticated AI models exhibit excellent reasoning, information retrieval, semantic analysis, and conversational abilities.

Unlike traditional chatbot systems which mostly utilize rule-based systems and static responses, generative AI

chatbots use deep learning and neural network methods such as transformers. This has greatly enhanced the ability of AI chatbots to adapt and reason in dynamic dialogues. Hence, these innovative AI models are now indispensable tools for academics and industries alike. Although there has been quick uptake of AI chatbots, assessment of their performance is still an uphill task owing to the multi-faceted human-AI interaction dynamics that are involved. Response accuracy, context appropriateness, usability, conversational ability, speed of interaction, reliability, and satisfaction among others are critical determinants of chatbot performance. Current literature has mostly addressed performance from a purely technical point of view using benchmarks or algorithm modifications, with scant research on comparative assessments in the presence of real-life interaction scenarios from a user perspective.

This research aims to conduct a comparative evaluation of generative AI chatbots through user-centered survey analytics and multi-dimensional performance

assessment techniques. This research focuses on the evaluation of user perception on different platforms of AI chatbot services regarding communication quality, context comprehension, semantic precision, efficiency of interaction, and overall satisfaction. A structured survey based on the involvement of 30 subjects was carried out to collect both quantitative and qualitative data on the performance of the chatbots under study. Some statistical techniques were used for the interpretation of user perceptions.

The significance of the current study lies in the fact that it introduces one more dimension into the expanding body of literature about evaluating conversational artificial intelligence systems. By analyzing feedback from users through a few evaluation criteria, the study provides an insight into the behavior of AI chatbots in real-world communication environments. This may help researchers, developers, teachers, and businesses select and develop better AI chatbot solutions according to their needs.

Literature review covering the use of generative AI chatbots and approaches to evaluating conversational AI is presented in Section 2. Section 3 includes information about the research methods, survey, and assessment criteria. The findings from the survey analysis and comparison of different solutions are presented in Section 4. The results and limitations of the current study are discussed in Section 5. The conclusion and directions for further research are provided in Section 6.

II. Related Works

Set The fast growth of generative Artificial Intelligence (AI) and large language models (LLMs) led to a significant interest in research on conversational AI systems and intelligent chatbots. Research papers explored the topic of conversational AI system performance and usability, as well as user experience and satisfaction with chatbots in a variety of application fields, such as education, healthcare, software engineering, customer service, and information retrieval.

According to the recent research paper published by Smutny and Bojko, an analysis of multiple generative AI chatbots was performed in relation to their abilities in helping software engineers program web pages and other objects [1]. The research indicated the ability of such chatbots to help developers perform their tasks.

Escobar-Grisales et al. developed an assessment model to quantify the efficiency of conversations between humans and AI chatbots through the application of deep learning methods with temporal neural networks [2]. It was important to consider conversational coherence, interaction consistency, and appropriate response when evaluating chatbots. The findings confirmed that the behavior of users could have a considerable effect on the effectiveness of conversations.

Regarding the field of health care, a comparative experimental study of the four most famous AI chatbots like ChatGPT, Bard, Bing AI Chat, and Claude AI were performed to evaluate their performance in responding to emergency situations in healthcare [3]. As a result, it was found out that LLM-based chatbots give proper responses to general medical questions, but the results may vary and become unreliable in some cases.

Also, Wangsa et al. have reviewed some early AI chatbots like ChatGPT, Bard, Llama, Ernie, and Grok from the perspective of education and healthcare [4]. According to the results, ChatGPT turned out to be superior to its rivals in writing assistance, educational counseling, and conversational capabilities. Nevertheless, it should be noted that other chatbots had their strengths in helping with multilingual conversation and specific subjects. Thus, one can claim that the selection of the chatbot depends on personal preferences, the sphere of usage, and evaluation criteria.

User-centered evaluation has gained much attention in research on conversational AI. Generative AI usability based on a large-scale analysis of user reviews gathered via mobile application platforms [5]. In the study under discussion, more than 11,000 reviews were assessed related to such generative AI applications as ChatGPT, Gemini AI, and Microsoft Copilot. According to the findings, usability, simple interface, fast responses, and contextual accuracy affected the user experience positively.

Another study investigated the factors affecting chatbot satisfaction through a mixed-method comparative approach in users from mainland China and Hong Kong [6]. It was found that the factors of information quality, empathy, availability, and response efficiency greatly affected user satisfaction and chatbot adoption intentions. This study also indicated that cultural and

regional differences could affect user perceptions toward conversational AI.

The Microsoft Research team came up with a novel framework to predict user satisfaction of conversation systems driven by LLMs [7]. Specifically, the SPUR framework relied on supervised prompting and rubric-based evaluations to assess user satisfaction in general-purpose and task-oriented chatbot systems. The study suggested that interpretable metrics were crucial to evaluate conversational quality of AI.

More recently, research has highlighted the inadequacy of benchmark-driven approaches for assessing the actual capabilities of chatbot systems. In a survey study conducted on a cross-platform basis with various AI chat assistants, it was found that users usually assess chatbot systems in terms of their interface usability, clarity of responses, contextual relevance, and utility [8]. This research further emphasized the presence of hallucination, faulty reasoning, and contextual mistakes in the current generative AI models.

A few studies have analyzed the constraints associated with generative AI chatbots. Findings obtained from studies conducted at Microsoft and Salesforce concluded that the accuracy of chatbots is compromised during lengthy discussions due to the accumulation of context and inconsistency in responses [9]. There were other studies that highlighted the issue of hallucinations and information dissemination caused by poor uncertainty handling in large language model architecture [10].

Despite the significant advances made in conversational AI technology, the present literature focuses mainly on benchmarking, domain-specific performance assessment, and architectural studies. Fewer studies examine user-based comparative analyses using real-life interactions with various chatbot systems. Therefore, this research aims to address this gap by conducting a survey-based comparative evaluation of generative AI chatbots using multi-dimensional performance assessment criteria including contextual understanding, semantic accuracy, usability, response efficiency, and user satisfaction.

III. Research Methodology

In this section, the methodology that is adopted in the comparative study of Generative AI Chatbots will be explained through user-based surveys and performance analysis using multi-dimensional techniques. The quantitative method of research will be applied to

conduct user perception and chatbot performance analysis in this study.

3.1 Research Design

The present study was carried out by using a comparative method where survey research was used for assessing modern Chatbot systems. The objective of the research was to analyze how the perceptions of the Chatbot are formed regarding such factors as conversational abilities, contextual awareness, semantics, usability, response time, and satisfaction from interaction. To reach this goal, a special questionnaire was developed.

The study follows four major stages:

1. Selection of AI chatbot platforms
2. Survey questionnaire design
3. Data collection from participants
4. Statistical and comparative performance analysis

The overall research framework is illustrated conceptually in table 1.

Table 1. Research Framework

Phase	Description
Data Collection	Survey responses from participants
Chatbot Interaction	Users interact with selected AI chatbots
Evaluation Criteria	Accuracy, usability, speed, context understanding
Statistical Analysis	Comparative evaluation and interpretation
Performance Assessment	Identification of strengths and limitations

3.2 Selected AI Chatbot Platforms

The study focuses on four widely used generative AI chatbot platforms:

- ChatGPT
- Gemini
- Claude
- Microsoft Copilot

The following chatbot systems were selected because of their popularity, accessibility, and advanced large language model capabilities. Each platform applies transformer-based architecture and generative AI techniques to support conversational interaction and intelligent response generation shown in table 2.

Table 2. Comparative Characteristics of Selected Chatbots

Chatbot	Developer	Key Strength
ChatGPT	OpenAI	Contextual understanding
Gemini	Google	Information retrieval
Claude	Anthropic	Conversational safety
Copilot	Microsoft	Productivity assistance

3.3 Survey Participants

A total of (30) participant was selected to be surveyed for this research. The selected participants were made up of university students, researchers, professionals, and users of chatbots prior to the experiment. The convenient sampling method was employed while selecting the participants because of their availability and awareness regarding the conversational AI technology. The inclusion of different demographic backgrounds among the participants was essential to get a broader perspective on chatbot performance.

3.4 Survey Questionnaire Design

The questionnaire was designed using a Likert-scale evaluation model ranging from 1 to 5 in the following table 3.

Table 3. Questionnaire design

Scale	Meaning
1	Poor
2	Fair
3	Average
4	Good
5	Excellent

The questionnaire consisted of multiple evaluation dimensions including:

- Response Accuracy
- Contextual Understanding
- Semantic Relevance
- Ease of Use
- Response Speed
- User Satisfaction
- Conversational Fluency
- Reliability of Information

Participants provided ratings based on their interaction experiences with each chatbot platform.

3.5 Data Collection Procedure

The research survey was conducted online through digital forms and direct participant interaction sessions. The selected participants were asked to test selected chatbot platforms using common conversational tasks such as:

- General question answering
- Educational assistance
- Content generation
- Problem-solving tasks
- Information retrieval

After interacting with the chatbots, participants completed the evaluation questionnaire. The data collected were organized and processed for statistical analysis.

3.6 Performance Evaluation Metrics

The comparative assessment of chatbot systems was performed using multiple performance dimensions. The overall chatbot performance score was calculated based on the average user ratings across evaluation criteria in this research.

3.7 Data Analysis Techniques

The survey data collected were analyzed using descriptive statistical techniques including:

- Mean analysis
- Percentage analysis
- Comparative ranking

- Standard deviation analysis

The results were presented using tables and comparative performance evaluations. Statistical analysis enabled the identification of strengths and weaknesses among the evaluated chatbot platforms. Comparative visualization techniques such as bar charts and radar charts can be applied to illustrate performance differences across evaluation dimensions.

3.8 Research Limitations

While the paper is an important source of information regarding the performance of generative AI chatbots, there are certain weaknesses in the work. To start with, the number of people who took part in the survey is only 30, which means that it does not have sufficient generalization. Secondly, participants used their subjective opinions in evaluating the performance of chatbots. Apart from the limitations mentioned above, this study makes an important contribution to ongoing research into the methodology of evaluating generative AI.

IV. Survey Analysis and Comparative Evaluation Results

Analysis and comparative evaluation results from the survey of generative AI chatbot platforms are provided in this section. Descriptive statistical techniques were used for the analysis of survey data gathered from the user-centric analysis of the performance of chatbots on various factors such as accuracy of answers, context awareness, conversation skills, usability, speed of responses, and user satisfaction.

4.1 Demographic Distribution of Participants

A total of 30 participants participated in the survey. The respondents included students, academic researchers, professionals, and general users with prior experience using AI chatbot systems. Table 4 and figure 1 presents the demographic distribution of participants.

Table 4. Participant Demographics

Category	Frequency	Percentage (%)
Male	18	60%
Female	12	40%
Students	14	46.7%
Professionals	9	30%
Researchers	4	13.3%
General Users	3	10%

The demographic analysis indicates that students and professionals represented the majority of chatbot users, reflecting the increasing adoption of generative AI technologies in educational and workplace environments.

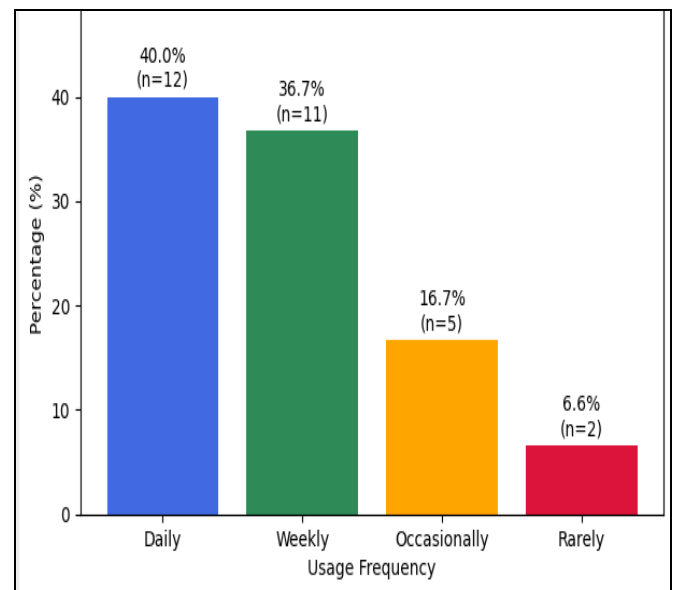


Figure 1. AI Chatbot Usage Frequency Among Participants

4.2 Chatbot Usage Frequency Analysis

Participants were asked about the frequency of their interactions with AI chatbot systems. The survey results are summarized in Table 5.

Table 5. Frequency of AI Chatbot Usage

Usage Frequency	Number of Participants	Percentage (%)
Daily	12	40%

Weekly	11	36.7%
Occasionally	5	16.7%
Rarely	2	6.6%

The results show that most participants regularly use AI chatbot systems for academic support, information retrieval, content generation, and communication assistance in figure 2.

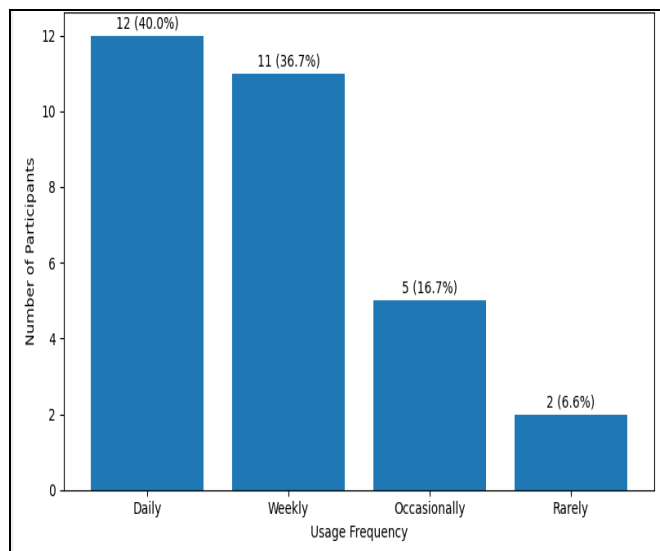


Figure 2. Usage Frequency of Participants

4.3 Comparative Performance Evaluation

The selected AI chatbot systems were comparatively evaluated based on user ratings across six major performance dimensions:

- Response Accuracy
- Contextual Understanding
- Conversational Fluency
- Ease of Use
- Response Speed
- User Satisfaction

Participants rated each chatbot using a 5-point Likert scale. The average scores are presented in Table 6 and figure 3.

Table 6. Comparative Performance Scores of AI Chatbots

Evaluation Criteria	Chat GPT	Gemini	Claude	Co pilot
Response Accuracy	4.7	4.3	4.4	4.1

Contextual Understanding	4.8	4.2	4.5	4.0
Conversational Fluency	4.7	4.3	4.4	4.1
Ease of Use	4.8	4.5	4.3	4.2
Response Speed	4.6	4.5	4.2	4.3
User Satisfaction	4.7	4.3	4.4	4.1

According to the comparative evaluation, it can be stated that ChatGPT proved to be the best among all evaluation parameters considered, specifically contextual knowledge, conversation fluency, and ease of use. Claude performed well in conversation coherence and safe interaction while Gemini was evenly matched in terms of information gathering and response time.

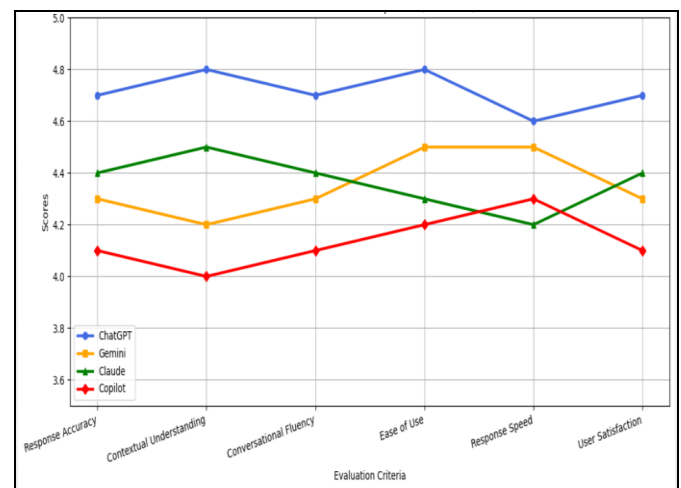


Figure 3. AI Chatbot Platforms Comparison

4.4 Mean Performance Analysis

The overall chatbot performance score was calculated using the mean evaluation formula:

$$\bar{X} = \frac{\sum X_i}{N} \quad (1)$$

Where,

- $\bar{X}$ represents the average performance score
- X_i represents individual participant ratings
- N represents the total number of participants

The calculated overall mean scores are shown in Table 7 and figure 4.

Table 7. Overall Mean Scores

Chatbot Platform	Mean Score
ChatGPT	4.72
Claude	4.37
Gemini	4.35
Copilot	4.13

The results indicate that ChatGPT obtained the highest overall mean score among the evaluated chatbot systems.

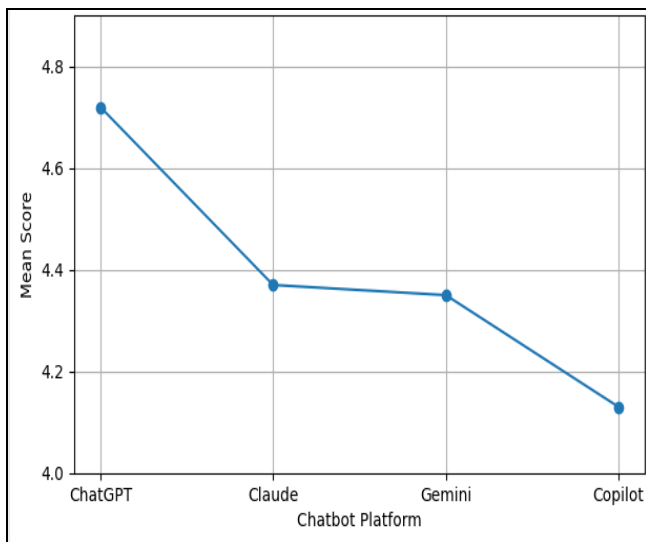


Figure 4. Mean Score Comparison of AI Chatbot Platforms

4.5 Standard Deviation Analysis

To evaluate response consistency among participants, standard deviation analysis was performed using the following equation:

$$\sigma = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N}} \quad (2)$$

Where,

- X_i = each data value
- $\bar{X}$ = population mean (average)
- N = total number of data values

The standard deviation results are presented in Table 8 and figure 5.

Table 8. Standard Deviation Scores

Chatbot Platform	Standard Deviation
ChatGPT	0.31
Claude	0.42
Gemini	0.45
Copilot	0.51

Lower standard deviation values indicate higher consistency in participant evaluations. ChatGPT demonstrated the most stable user satisfaction levels among the evaluated platforms.

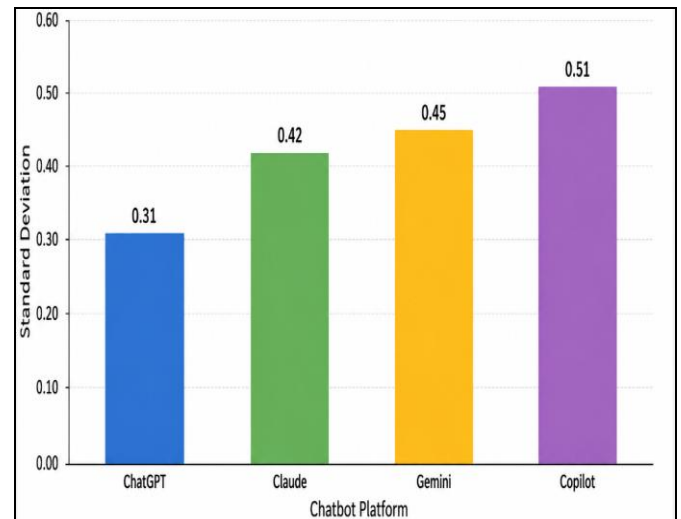


Figure 5. Standard Deviation of Chatbot Platforms

4.6 User Satisfaction Analysis

Participants were additionally asked whether AI chatbots improved productivity and communication efficiency shown in table 9 and figure 6, 7.

Table 9. User Satisfaction Responses

Response Category	Frequency	Percentage (%)
Strongly Agree	13	43.3%
Agree	11	36.7%
Neutral	4	13.3%
Disagree	2	6.7%

Strongly Disagree	0	0%
-------------------	---	----

The findings reveal that approximately 80% of participants positively perceived AI chatbots as effective tools for improving productivity, communication, and information access.

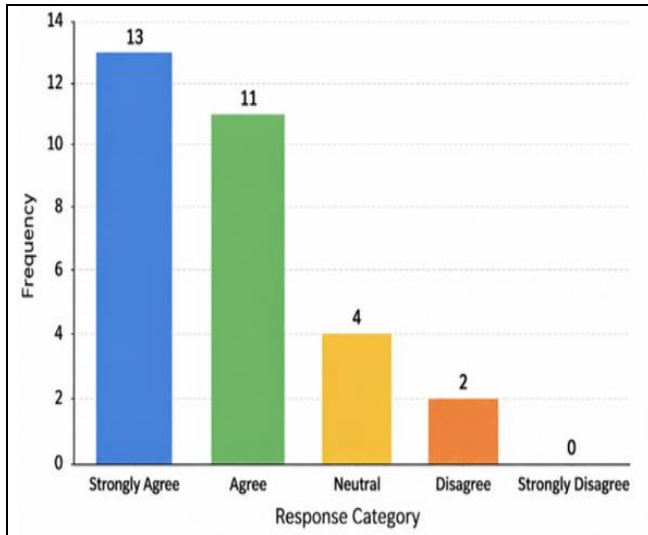


Figure 6. Response Category Distribution

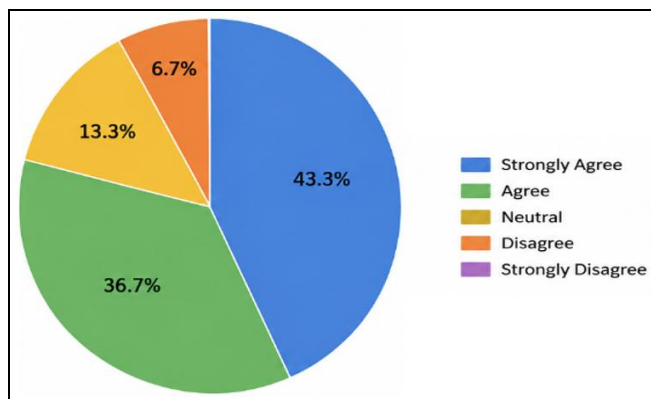


Figure 7. User Satisfaction Percentage

4.7 Comparative Findings

According to the analysis of the surveys conducted, generative AI chatbots are very intelligent, as compared to rule-based chatbots, and provide high-quality interactions. Out of all the chatbots analyzed, ChatGPT was the best due to high-quality contextual awareness and coherence, as well as its friendly conversational tone.

Stability and safety of conversation were qualities attributed to Claude, while effective collection of information and response generation were qualities attributed to Gemini. The productivity and efficiency of integration were observed in Microsoft Copilot; however, the conversational intelligence was lacking.

Despite all the advantages associated with the chatbots, however, there were some disadvantages of generative AI chatbots identified by the participants' responses on the survey. Some of these disadvantages are the existence of misinformation, inconsistency in reasoning in complex situations, repetition, and lack of expertise in a particular field. From the above discussion, it is evident that user-centered evaluation methods play an important role in evaluating chatbot efficiency.

V. Discussion

Fast development of transformer architecture allowed us to enhance the quality of human-computer interaction and to make chatbots give more realistic, relevant and consistent with context answers. As can be seen from the results of the survey, users regard generative AI chatbots as effective means of communication with others, finding information, studying and enhancing productivity. One of the major findings of the current study is the high effectiveness of the most sophisticated AI chatbot generators when it comes to their capacity to comprehend context and ensure conversation continuity. The platform that performed the best in all respects was ChatGPT. Based on the feedback provided by the respondents, ChatGPT gave the most coherent and relevant answers, showed greater conversation persistence, and understood complex instructions better than any other platform.

In addition, it was found out that Claude was characterized by high conversational consistency and numerous other characteristics associated with safety. In their opinion, Claude was a safe and consistent partner who could be trusted in discussing complicated matters and accomplishing serious professional tasks. This finding proves the growing popularity of conversational AI regarding the issues related to the development process and associated safety parameters. At the same time, Gemini was also characterized by high results in terms of information retrieval and response time. As such, it can be viewed as an efficient tool for processing requests and providing access to knowledge base information.

Another significant finding obtained from this research is the growing dependence of users on chatbot assistance in accomplishing various academic and professional tasks. All participants reported that they used chatbot assistance consistently in writing, idea generation, studying, coding, and other problem-solving activities.

Despite acknowledging the presence of positive feedback concerning the results achieved, some issues regarding the current models of generative AI were also discovered. It was noted that hallucinations, creation of wrong information, repetitions, and logical inconsistencies are among the main disadvantages of existing AI. Hallucinations represent one of the major disadvantages of large language models, which might result in the fact that chatbots will generate wrong information. The disadvantages mentioned might affect the users' satisfaction when it comes to sensitive issues such as healthcare, finance, or legal advice.

Additionally, it became apparent from the survey that users not only expect accurate information from the chatbots but also want them to be convenient to use and fast. Particularly, users preferred interacting with bots whose behavior resembled human interactions, and which were fast.

One more important aspect highlighted by the research is that the evaluation of chatbots is multi-dimensional. It cannot be measured with one metric only because human interaction assumes understanding of meaning, emotion, context, linguistic skills, and relevance of messages exchanged. Thus, using various metrics may help to gain a clearer picture of chatbot performance.

Limitations of this research include the limited number of participants involved in the experiment (30 in total). In general, the analysis highlights the significant advancements made by conversational AI chatbots in terms of intelligent conversation and user interactions. Nevertheless, issues concerning reliability, consistent reasoning, and accuracy of facts need to be addressed in future studies and technology developments. This research proves the importance of integrating survey analytics based on users' perceptions along with performance evaluation approaches for assessing contemporary conversational AI systems.

VI. Conclusion and Future Work

In this paper, there has been a comparative analysis of some popular generative AI chatbot applications through user-centric survey analytics as well as multiple-performance measurement approaches. The main goal is to understand and evaluate the effectiveness of popular chatbot applications based on factors such as response accuracy, context recognition capabilities, fluency of interaction, ease-of-use, response time, and customer satisfaction. With the use of survey data from 30 users, this paper offers a

systematic review of modern conversational AI applications.

ChatGPT was the most effective of all the selected platforms in various dimensions like contextual knowledge, conversational ability, and satisfaction rate. Claude scored high in safe conversational behavior, while Gemini excelled in information search capabilities and conversation performance. The Microsoft Copilot performed very well when it came to productivity-based conversations but had relatively lower scores in conversational performance.

This research also shows that user satisfaction rates depend on various factors other than accuracy, such as usability, answer clarity, context awareness, and interaction quality. Most users perceived the usage of AI chatbots positively for educational and research purposes as well as their involvement in work. Nonetheless, problems like hallucination, reasoning inconsistencies, and false information are among the most prominent concerns regarding their utilization. In conclusion, this study shows the importance of considering the user-focused, multidimensional assessment methodology to evaluate the success of generative AI chatbots properly.

Although this paper provides much insight into the topic of chatbot performance comparison, there are still some points that might be further investigated. Firstly, future research papers can consider the possibility of increasing the number of subjects and collecting data from people who have various social status and belong to different professions. This would increase the reliability and significance of the results obtained.

Secondly, along with the subjective information received from the survey, the researchers can add the information related to the behavior of the participants and gathered in the form of logs of interactions, response time statistics or even percentage of tasks done. Thirdly, more advanced methods of data analysis, like regression analysis, factor analysis or even SEM, can be applied to establish relationships and dependencies between the evaluation criteria and user satisfaction. Fourthly, given the growing popularity of chatbots, future research can investigate their performance comparison within a particular area, like healthcare, educational technologies, financial management or customer service.

Finally, since the area of generative AI is developing very rapidly, it becomes necessary to constantly

monitor the performance of chatbots. The chatbot performance comparison framework will have to adapt accordingly as the AI models improve. In conclusion, this paper has contributed to the new field of conversation AI evaluation through an empirical framework for chatbot performance comparison. This research has provided meaningful lessons that will be helpful to scholars, practitioners, and institutions seeking to enhance the development of intelligent conversation platforms.

and effectiveness. *Indian journal of science and research*, 3(3), 88-93.

References

- [1] Smutny, P., & Bojko, M. (2024). Comparative analysis of chatbots using large language models for web development tasks. *Applied Sciences*, 14(21), 10048.
- [2] Escobar-Grisales, D., Vásquez-Correa, J. C., & Orozco-Arroyave, J. R. (2024). Evaluation of effectiveness in conversations between humans and chatbots. *Multimedia Tools and Applications*.
- [3] Sallam, M. (2023). ChatGPT utility in healthcare: A comparative evaluation of AI chatbot responses. *Journal of Medical Internet Research*, 25.
- [4] Wangsa, K., Karim, S., Gide, E., & Elkhodr, M. (2024). A systematic review and comprehensive analysis of pioneering AI chatbot models. *Future Internet*, 16(7), 219.
- [5] Alabduljabbar, R. (2024). User-centric AI: Evaluating the usability of generative AI applications through user reviews on app stores. *ResearchGate Preprint*.
- [6] Xu, Y., & Liu, Y. (2023). Can chatbots satisfy me? A mixed-method comparative study of satisfaction with task-oriented chatbots. *Computers in Human Behavior*, 145.
- [7] Lin, Z., et al. (2024). Interpretable user satisfaction estimation for conversational systems with large language models. *Microsoft Research Technical Report*.
- [8] Zhang, H., & Chen, L. (2025). Beyond benchmarks: How users evaluate AI chat assistants. *arXiv preprint arXiv:2603.25220*.
- [9] Microsoft Research & Salesforce (2025). Analysis of conversational AI reliability and long-context performance degradation. *AI Systems Report*.
- [10] OpenAI. (2025). Understanding hallucinations in large language models. *OpenAI Research Blog*.
- [11] Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.
- [12] Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS*.
- [13] OpenAI. (2023). GPT-4 technical report. *OpenAI Technical Report*.
- [14] Google DeepMind. (2023). Gemini technical overview. *Google AI Research Documentation*.
- [15] Anthropic. (2023). Claude: Constitutional AI and harmlessness training. *Anthropic Technical Report*.
- [16] Microsoft. (2023). Copilot: AI-powered productivity assistant overview. *Microsoft Official Documentation*.
- [17] Thant, K. S., & Tin, H. H. K. (2023). The impact of manual and automatic testing on software testing efficiency