Journal of Database Management
Volume 35 • Issue 1 • January-December 2024

Identifying Alternative Options for Chatbots With Multi-Criteria Decision-Making:
A Comparative Study

Praveen Ranjan Srivastava
https://orcid.org/0000-0001-7467-5500
Indian Institute of Management, Rohtak, India

Harshit Kumar Singh
https://orcid.org/0000-0002-4107-8387
Indian Institute of Management, Rohtak, India

Surabhi Sakshi
Indian Institute of Management, Rohtak, India

Justin Zuopeng Zhang
https://orcid.org/0000-0002-4074-9505
University of North Florida, USA

Qiuzheng Li
Zhejiang Wanli University, China

ABSTRACT

Artificial intelligence-powered chatbot usage continues to grow worldwide, and there is ongoing research to identify features that maximize the utility of chatbots. This study uses the multi-criteria decision-making (MCDM) method to find the best available alternative chatbot for task completion. We identify chatbot evaluation criteria from literature followed by inputs from experts using the Delphi method. We apply CRITIC to evaluate the relative importance of the specified criteria. Finally, we list popular alternatives of chatbots and features offered and apply WASPAS and EDAS techniques to rank the available alternatives. The alternatives explored in this study include YOU, ChatGPT, PerplexityAI, ChatSonic, and CharacterAI. Both methods yield identical results in ranking, with ChatGPT emerging as the most preferred alternative based on the criteria identified.

KEYWORDS

Chatbots, CRITIC, Decision-Making, EDAS, MCDM, WASPAS

Chatbots have experienced renewed interest in recent years due to the increased capabilities of artificial intelligence (AI)-based conversational tools. Advancements in technology have directed a dramatic shift in virtual communication, particularly in automated language-related task completion. These advancements have led to more practical applications of such services (Peng & Bhaskar, 2023). Among the world's 7.3 billion people in 2015, 6.1 billion had a mobile phone that could send and receive short message service (SMS) messages, while Facebook alone had more than 1 billion members (Dale, 2016). Users of such platforms are now used to having text-based exchanges with one another online compared to voice-based conversations. The shift in consumer behavior makes chatbots an attractive option for many businesses. Companies worldwide have responded to the rising demand for chatbots by creating various resources and platforms that make the technology accessible

DOI: 10.4018/JDM.345917

to a broad audience. The chatbot-building tools market is estimated to have expanded 29% between 2018 and 2020 (Meisel, 2021). With fewer customers needing to contact a person, firms do not have to hire more people to keep up with demand or maintain a 24-hour support team. Additionally, chatbots may aid organizations in providing individualized care to each customer.

Chatbots are an example of cutting-edge technology that may seem elementary today. Twenty-three percent of companies providing customer service say they utilize chatbots powered by artificial intelligence . Despite this, more than three-quarters of consumers expect companies to implement improved technology to enhance services (Bitner et al., 2022). Users may pose queries, make assertions, or issue directives to these conversational AIs. One such emerging example of an AI-based chatbot is ChatGTP. With OpenAI's ChatGPT model, users may pose queries that are then replied to by an AI taught via supervised, reinforced machine learning. The replies are contingent on the information the algorithm has been fed by its users. As an acronym for "generative pretrained transformer," ChatGPT describes a chatbot with conversational capabilities. It leverages supervised and reinforcement learning to train the model, ranking the machine's replies as positively as possible (Mahesh, 2020).

Selecting the right platform may be a time-consuming process since there are a plethora of options out there. Companies or individuals looking to boost consumer interaction, broaden their client base, and create more leads should familiarize themselves with the characteristics of a chatbot. It is not only the abundance of options for chatbot development platforms and technologies that makes choosing one challenging; the choice also relies heavily on the problem domain that the chatbot will address. Features of the chatbot platform play a crucial role in predicting its effectiveness. This research addresses this issue by outlining a set of criteria businesses may use to choose an alternative. A chatbot criterion might be any one of its features or any other property. The result of the current study can be used by organizations to identify the alternatives relevant to their use case based on the criteria specified in this study. The relative weights of these criteria from the user's perspective give further insight into their importance. A combination of methods from the literature was employed to help us arrive at this set of standards. We conduct a literature review drawing from previously published works and recent research. We use Rogers's diffusion of innovation theory to evaluate users' perceptions of different chatbot characteristics and identify critical platform criteria.

Using this approach requires weighing the significance of each criterion in light of the chatbot's use case and evaluating how effectively the platforms under consideration support these criteria. We chose five widely used language-related AI platforms to test our suggested strategy for platform selection. There is a dearth of literature on chatbot development and assessing such platforms. The influence of chatbots in various applications has been the primary focus of the existing literature. In one such attempt, researchers examined whether a chatbot designed to mimic a particular celebrity might enhance the connection between distant learners and course materials. The number of factors included in other research that provides a mechanism for choosing a platform based on company characteristics is limited (Gulum et al., 2021). The comparisons required rise exponentially with factors included, making it exceedingly time-consuming and inconvenient even if they considered adding additional features to their assessment technique. We propose a different approach that works better on larger scales to address this issue. This would allow for a more thorough evaluation of available options, which might lead to a more reliable platform recommendation.

In the next section, we discuss the current status and the findings from the existing literature about chatbots, followed by a discussion of the methodology. Subsequently, the proposed model is detailed with results and sensitivity analysis results. Toward the end, we present implications and future research directions before concluding.

LITERATURE REVIEW

Studies Related to the Evaluation of Chatbot

Evaluation of chatbots can be approached from various perspectives, including user satisfaction, task completion rate, conversational quality, and system performance (Cai et al., 2022). Several studies have investigated the evaluation of chatbots from these perspectives. For instance, in a survey by Mokmin and Ibrahim (2021), the authors evaluated the performance of a chatbot designed to provide education on health support to college students. The study found that the chatbot helped 73.3% of responders grasp health concerns and have pleasant conversations. Similarly, in a study, the authors evaluated the effectiveness of a chatbot designed to help tourists. The study found that user satisfaction depends on the chatbot's informativeness, empathy, and interactivity (Orden-Mejía & Huertas, 2022).

Other studies have focused on the conversational quality of chatbots. For example, Barletta et al. (2023) evaluated a chatbot's ability to converse with users informally. Similarly, in another study, the authors assessed the conversational quality of chatbots using a human evaluation metric. The study found that the chatbot could generate responses that were rated similar to those generated by humans, indicating that the chatbot could maintain natural and engaging conversations with users.

Despite these promising results, the evaluation of chatbots is not without its challenges. One of the significant challenges is the lack of standardized evaluation metrics. Currently, there is no universally accepted set of metrics for evaluating chatbots, which makes it difficult to compare the performance of different chatbots. Liang and Li (2021) provided the solution through standard criteria and definitions for chatbot evaluation. Another challenge is the lack of diverse data sets for training and testing chatbots. Most existing data sets focus on specific domains, such as customer service or restaurant booking, which limits the generalizability of chatbots to other domains (Narducci et al., 2020).

In conclusion, evaluating chatbots is a complex and challenging task requiring a multifaceted approach. While several studies have demonstrated the effectiveness of chatbots in various domains, standardized evaluation metrics and more diverse data sets still need to be standardized. Computers that performed well on the measures used to evaluate the results of natural language processing AI do not match user expectations. This points to a gap in the comprehensiveness of the existing evaluation metrics. Nonetheless, few research studies have examined how various factors affect user experience on chatbot platforms. Table 1 presents a list of such studies. This research aims to quantify the factors associated with chatbot utility that are most important to the users.

The assessment of platform utility in the context of chatbots has not yet been carried out, as shown in Table 1. Previous research has looked at how chatbots influence users but not at how various aspects of a platform's experience stack up against one another. Experimentation and empirical analysis using survey approaches dominate the bulk of the investigations. Figure 1 shows that there is still a need for research on how to evaluate platforms based on the utility of chatbots.

Multicriteria decision-making (MCDM) is a robust method for evaluating chatbots that fills this void by enabling decision-makers to consider multiple criteria simultaneously (Yalcin et al., 2022). This is especially crucial when judging chatbots, which are typically built to accommodate various users and accomplish numerous objectives.

MCDM

User satisfaction, conversational quality, task completion rate, and overall system performance can be measured with MCDM's organized framework. MCDM allows decision-makers to prioritize different criteria and use mathematical models to compare how well various alternatives perform across those criteria. With MCDM, users can evaluate chatbots with thorough and dependable criteria. This method gives decision-makers a complete picture of the pros and cons of available chatbots, allowing them to make better choices. Using MCDM methods, decision-makers can weigh the relative importance of competing criteria. A chatbot's strengths may lie in one area, while its weaknesses lie

Figure 1. Research Gaps

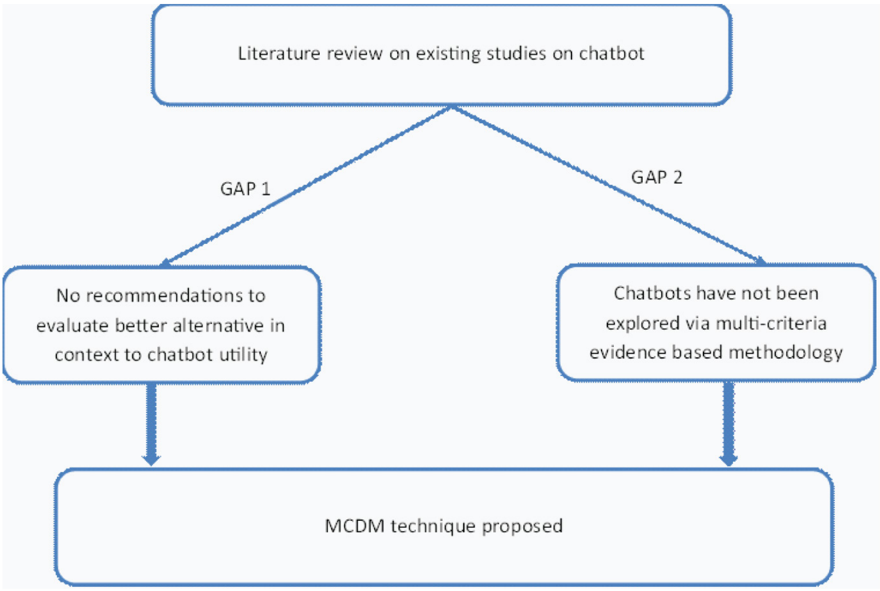

Literature review on existing studies on chatbot
GAP 1
GAP 2
No recommendations to evaluate better alternative in context to chatbot utility
Chatbots have not been explored via multi-criteria evidence based methodology
MCDM technique proposed

Table 1. Existing Literature on Chatbot's Utility Criteria

Reference	Method of evaluation	Domain	Objective
Orden-Mejía & Huertas, 2022	Exploratory factor analysis; Hierarchical regression	Tourism	Investigation of factors contributing to user satisfaction with chatbots.
Barletta et al., 2023	Multicriteria decision-making (MCDM) (AHP)	Healthcare	Assess the quality of the medical chatbot and compare a chatbot's two different iterations.
Cai et al., 2022	Task-oriented user studies	Music	Investigate the effectiveness of Dialogue-based conversational recommender systems.
Mokmin and Ibrahim, 2021	Mixed method study	Education	Analysis of the efficacy, performance, and technological adoption of a chatbot created to educate users and provide health literacy.
Narducci et al., 2020	Experimental evaluation	Music	Analysis of how various interaction styles affect recommendation precision and user input costs
Sugisaki and Bleiker, 2020	Research-based method	Language interaction	Synthesis of the linguistic concepts necessary for a discussion in a natural language
Huang et al., 2018	Survey	General knowledge, reasoning, memory, and personality.	Assess the capabilities of Tarie, a conversational AI.
AbuShawar and Atwell, 2016	Quantitative and Qualitative evaluation	Language interaction	The paper discusses black box, glass box, comparative, quantitative, and qualitative natural language conversation system assessment methods.
Liang and Li, 2021	Review	Criteria and definitions	Provide standard criteria and definitions for chatbot evaluation

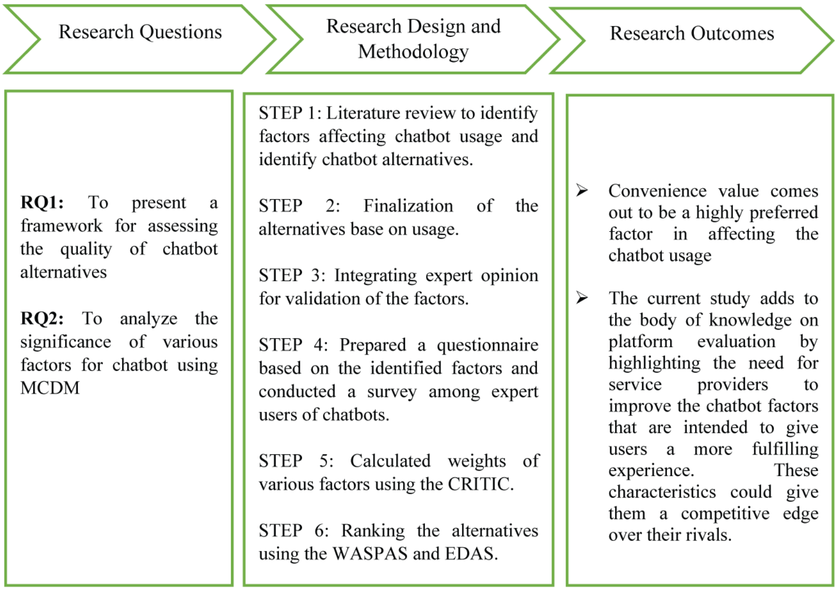
Table 2. Studies Applying MCDM to Evaluate Traditional Chatbots' Utility

Reference	Factors	Objective
Chakrabortty el al. (2023)	Security, Speed, Responsiveness, Satisfaction, Reliability, Assurance, Tangibility, Engagement, And Empathy	Chatbot selection in the telecommunication industry
Singh et al., 2022	Effectiveness, Speed, Safety, Ease of Use	Vital chatbot factors for Banking.
Syamsuddin and Warastuti, 2021	Reliability, Usability, Efficiency, Maintainability, Portability	ChatBot Platform for Health Enterprise Training
Santa et al., 2023	Effectiveness, Efficiency, Satisfaction, Risk, Context completeness, Flexibility	Measure chatbot quality
Hsu, 2023	Technologies, Goals, Boundaries, Activities	Introducing Chatbots into Mental Health Services
Phooriyaphan and Rachsiriwatcharabul, 2022	Speed, Flexibility, Accuracy, User friendly, Privacy, Functionality, Enjoyment, Security and privacy, Assurance, Design, Convenience, Customisation	Healthcare Chatbot Evaluation

in another. The theory of decision science can aid in weighing the significance of each criterion and striking a happy medium. Methods from MCDM can be used to evaluate chatbots against the market's other options. Innovative chatbots that stand out in the market and provide a competitive advantage can be found by decision-makers by considering various factors.

The extant literature has no research on the application of MCDM on AI-powered chatbots. Table 2 summarizes the previous studies that employ MCDM to study traditional chatbots. The current study goes beyond the existing research by incorporating factors that become relevant for AI-powered chatbots. We classify the factors identified by earlier studies under appropriate categories. While the existing papers cover the chatbots' user satisfaction and technology aspects, factors such as linguistic

Figure 2. Framework of the Paper



RQ1: To present a framework for assessing the quality of chatbot alternatives

RQ2: To analyze the significance of various factors for chatbot using MCDM

STEP 1: Literature review to identify factors affecting chatbot usage and identify chatbot alternatives.

STEP 2: Finalization of the alternatives base on usage.

STEP 3: Integrating expert opinion for validation of the factors.

STEP 4: Prepared a questionnaire based on the identified factors and conducted a survey among expert users of chatbots.

STEP 5: Calculated weights of various factors using the CRITIC.

STEP 6: Ranking the alternatives using the WASPAS and EDAS.

- Convenience value comes out to be a highly preferred factor in affecting the chatbot usage
- The current study adds to the body of knowledge on platform evaluation by highlighting the need for service providers to improve the chatbot factors that are intended to give users a more fulfilling experience. These characteristics could give them a competitive edge over their rivals.

or information retrieval capabilities are missing from the analysis. The current study also incorporates these factors through a literature review comparing chatbot features with expert opinions.

Decision-Science Theory

Stakeholder evaluation of decommissioning outcomes and decisions may be significantly influenced by how they rank, balance, and prioritize features. Stakeholders' values might differ depending on several factors, including religious convictions, individual interests, and prior experiences. Decision-science theory expands on different stakeholder values. According to Keeney (1992), the ideal choice is the one that most accurately reflects the stakeholders' values. It implies that a stakeholder's decision to decommission a system will be more or less acceptable depending on their values in a specific context. For instance, the adopted multicriteria decision analysis tools' criteria and weighting should precisely reflect the actual stakeholder values to provide outcomes and conclusions more likely to be accepted by stakeholders (Tung, 2021). Consumers behave differently with different media characteristics, as each character provides unique value. Hence, it is imperative to understand the characteristics of the media itself.

Since chatbots are meant to help users make decisions, and decision science provides a framework for understanding how humans make decisions, the two fields can complement each other in evaluating chatbots. In particular, decision-science theory can be used to assess how well chatbots help with decision-making and where they could be improved. Decision-science theory, for instance, suggests that people are more likely to make good decisions when they have access to all relevant information, that the data are presented clearly and understandably, and that they can weigh the pros and cons of different options. The effectiveness of chatbots accommodating these aspects of the decision-making process can be measured. Feedback and iterative decision-making are crucial components of decision-science theory. Chatbots can be designed to provide feedback to users on their decision-making process and learn from user interactions to improve their decision-making recommendations over time.

Overall, decision-science theory provides a framework for evaluating chatbots as decision-making tools and designing chatbots that better support users' decision-making processes.

Diffusion of Innovation Theory

According to diffusion of innovation theory, the likelihood of an invention's widespread adoption may be predicted partly by looking at how its attributes connect to prospective users' adoption and usage behaviors (Rogers, 1995). Rogers defines the five distinguishing features of innovations. These are relative benefit, compatibility, complexity, trialability, and observability of innovations. Rogers's research has been cited by experts from many fields who have studied the correlation between innovative traits and subsequent uptake and use.

An invention has a relative advantage if it is seen as better than, superior to, or beneficial in some way compared to the concept that came before it. The complexity of an invention is measured by how difficult it is to grasp and apply (Chwelos et al., 2001).

The degree to which a particular group of people approves of innovation is proportional to the degree to which they believe it to be congruent with that group's values, experiences, and requirements. In a nutshell, the propensity to accept and use innovation is theorized to be connected to factors such as a high relative benefit, low complexity, and high compatibility.

Chatbots can be evaluated with the help of diffusion of innovation theory because they provide a framework for analyzing the introduction, adoption, and spread of novel technologies. The commercial success of chatbots can be measured, in particular, by employing diffusion of innovation theory. According to diffusion of innovation theory, several steps involve getting people to try and eventually adopt new technologies. When applied to chatbots, this theory suggests that widespread adoption depends on the chatbot's ability to advance through these stages. Potential users weigh the pros and cons of new technology during this phase of the innovation diffusion process, making evaluation a crucial step. Evaluating a chatbot could entail rating its prowess in understanding user intent, information provision, and conversational naturalness. Developers and marketers can benefit from a more nuanced understanding of what drives chatbot success by evaluating chatbots through the lens of diffusion of innovation theory. With this knowledge, they can create chatbots that appeal to a broader audience and increase engagement, leading to better business outcomes.

Attributes of Chatbots

By stimulating interest, tailoring answers to individual needs, and providing rich behavioral data, chatbots have the potential to transform the conversational experience completely. Conversations that the user leads have the potential to depart from the chatbot's script since AI powers them. As discussed earlier, Table 2 presents the criteria used in the literature for chatbot evaluation. We extend the list by classifying the requirements based on a framework discussed in the next section and incorporating additional criteria absent from the earlier studies.

Russell-Rose (2017) proposed four broad perspectives for chatbot evaluation. These included the user experience perspective, information retrieval perspective, linguistic perspective, and technology perspective. User experience focuses on how the users of the chatbot relate to its navigability, assistance, and privacy. Information retrieval focuses on the chatbot's ability to detect intent and appropriately respond with adequate information. Linguistic performance measures the relevancy and unambiguity of the responses and their connection to the overall theme of the conversations. Finally, the technology perspective focuses on the chatbot's ability to learn to understand and its response time frequency.

We grouped the chatbot traits observed in the literature along these criteria. This approach allows us to gather a thorough and detailed understanding of user experiences and preferences. The different categories of these perspectives have been carefully chosen to encompass the various aspects of human-chatbot interactions. We identified a questionnaire for the criteria categories as our initial list of evaluation metrics.

RESEARCH METHODOLOGY

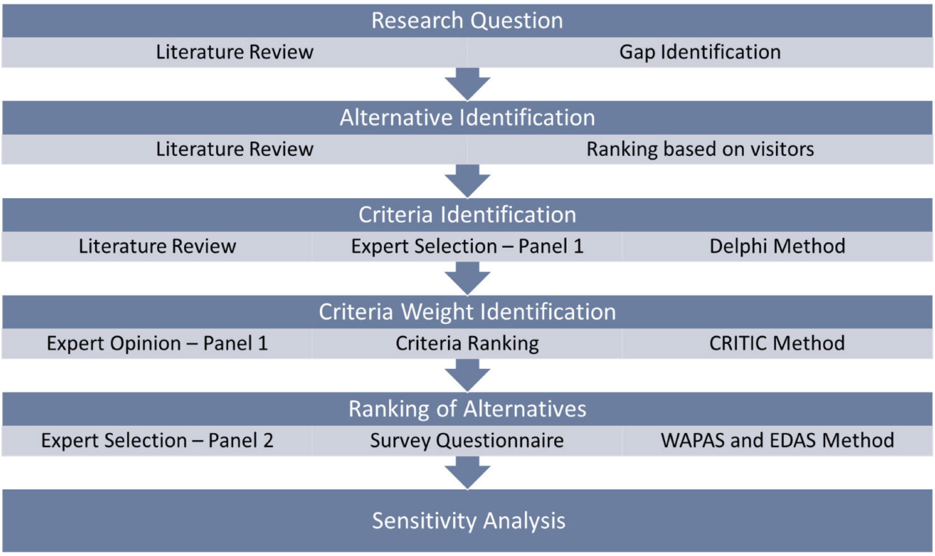
The employed model considers multiple factors to enhance the chatbot's overall success. To ensure the reliability of the findings, we applied the criteria importance through intercriteria correlation (CRITIC) method to determine the relative weights of the various evaluation factors under consideration. This approach proves more suitable than other MCDM techniques, as the study's primary focus lies in assessing the relative importance of one factor compared to another. It also strives to gain insights into the consumer's viewpoint regarding the factors contributing to platform quality improvement (Rokhsaritalemi et al., 2022). Subsequently, the performance of the chatbots was evaluated based on the criteria specified, and weighted aggregated sum product assessment (WASPAS) (Yel et al., 2021) was applied to rank the platform in order of performance. The evaluation based on distance from average solution (EDAS) technique was further applied to confirm the ranking result. We evaluate five distinct chatbot alternatives.

The research methodology is detailed in Figure 3. The attributes of chatbots are identified by extracting factors from the literature and improving them by incorporating insights from experts motivated by innovation theory. Further, drawing from decision-science theory, we use MCDM to identify the importance of factors and subsequently rank the identified alternatives.

CRITIC

The relative importance placed on the various features of an alternative during the selection process is crucial. In the literature on MCDM, multiple methods for estimating the relative importance of criteria are discussed. The CRITIC technique is one of the ways through which objective weighting can be accomplished (Diakoulaki et al., 1995). The correlations between the many criteria make up the most critical aspect of the methodology. More specifically, the criteria weights derived from

Figure 3. Research Methodology



the correlation analysis are combined with the contrast intensities calculated based on the standard deviations of the criteria (Jati et al., 2021).

We began by constructing the decision matrix X. It shows how the performance of various alternatives varies depending on several characteristics. The criteria (the objectives) and the possibilities are listed in the decision matrix's columns and rows, respectively, as shown in equation (1).

$$\mathrm{X} = [\mathrm{x}_{\mathrm{ij}}]_{\mathrm{mxn}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{1}$$

where $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, and x_{ij} presents the rating of the i_{th} alternative on the j_{th} criterion.

The second step is to normalize the decision matrix, as shown in equation (2).

$$x^*_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \tag{2}$$

where $\mathrm{x}^*_{\mathrm{ij}}$ represents the normalized value for x_{ij}.

Next, the criteria weights are established by considering each criterion's standard deviation. The weight of the j^{th} criterion, abbreviated w_{j}, can be calculated as shown in equation (3).

$$w_j = \frac{C_j}{\sum_{j=1}^{n} C_j} \tag{3}$$

where:

$$C_j = \sigma_j \sum_{j=1}^{n} \left(1 - r_{jj'}\right)$$

WASPAS

WASPAS is based on a combination of the weighted sum model (WSM) and the weighted product model (WPM) (Yucenur, 2021). To use WASPAS, the decision matrix components need first to be linearly normalized, as shown in equation (4).

$$\bar{x}_{ij} = \frac{x_{ij}}{max_i\, x_{ij}} \tag{4}$$

where x_{ij} is normalized.

The WSM technique is a weighted mean success criterion. It is an MCDM strategy used to analyze several possibilities in light of various selection criteria. Using the WSM technique in the manner outlined, we can ascertain the overall relative importance of the i^{th} choice and its place in the rankings (Triantaphyllou & Mann, 1989). If w_j is the weight of the jth criteria, the formula for relative importance is shown in equation (5).

$$Q_i^{(1)} = \sum_{j=1}^{n} \bar{x}_{ij} w_j \tag{5}$$

For the WPM technique, the formula to determine the total relative value for the alternatives is shown in equation (6).

$$Q_i^{(2)} = \prod_{j=1}^{n} (\bar{x}_{ij})^{wj} \tag{6}$$

Equation (7) is an example of a proposed joint generalized criterion for the aggregate of the two techniques (Zavadskas et al., 2013).

$$Q_i = 0.5\ Q_i^{(1)} + 0.5\, Q_i^{(2)} = 0.5 \sum_{j=1}^{n} \bar{x}_{ij} w_j + 0.5 \prod_{j=1}^{n} (\bar{x}_{ij})^{wj} \tag{7}$$

The alternatives are now ranked according to their Q values. Hence, the option with the highest Q value would be the best.

EDAS

When there is a conflict between the criteria, an innovative method of MCDM called EDAS is utilized. In EDAS, the optimal alternative is determined by calculating the distance from the solution that is most common. In EDAS, the positive distance from average (PDA) and the negative distance from average (NDA) are the two metrics that are utilized to assess the relative importance of the various options. These metrics illustrate the degree of dissimilarity between each alternative answer and the mean response. An alternative method that performs better than the standard technique is denoted by higher values of PDA and lower values of NDA in the solution's evaluation. The methodology presented by Huang et al. (2021) was utilized in this investigation to compute the ranking of vocations, with “m” options and “n” criteria being taken into consideration. Next, we describe the steps used in EDAS.

In step one, a decision matrix is formed, followed by taking the average solution (AV) of all criteria, as shown in equation (8).

$$AV = \left[A\,V_j\right]_{lxn} \tag{8}$$

where:

$$AV_j = \frac{\sum_{i=1}^{m} x_{ij}}{m}$$

In equations (9) and (10), we calculate the PDA and NDA matrices in the third step by considering all the criteria. These matrices highlight the disparity between the alternative solution and the average solution.

$$PDA = \left[PDA_{ij}\right]_{mxn} \tag{9}$$

$$NDA = \left[NDA_{ij}\right]_{mxn} \tag{10}$$

If jth criterion is beneficial, then one must use equations (11) and (12).

$$PDA_{ij} = \frac{\max\left(0, \left(x_{ij} - AV_j\right)\right)}{AV_j} \tag{11}$$

$$NDA_{ij} = \frac{\max\left(0, \left(AV_j - x_{ij}\right)\right)}{AV_j} \tag{12}$$

Then, in the fourth step, as shown in equations (13) and (14), the weighted sum of PDA (WP_i) and NDA (WN_i) for all alternatives is calculated.

$$WP_i = \sum_{j=1}^{n} w_j PDA_{ij} \tag{13}$$

$$WN_i = \sum_{j=1}^{n} w_j NDA_{ij} \tag{14}$$

where w_j is the weight of jth criterion. In this study, w_j is derived from the CRITIC method.

Next, we normalized the values, as shown in equations (15) and (16).

$$NWP_i = \frac{WP_i}{max_i\left(WP_i\right)} \tag{15}$$

$$NWN_i = 1 - \frac{WN_i}{max_i\left(WN_i\right)} \tag{16}$$

All alternatives' appraisal scores (AS) are calculated in the last step, as shown in equation (17).

$$AS_i = \frac{1}{2}\left(NWP_i + NWN_i\right) \tag{17}$$

Table 3. Expert Panel 1

Designation	Background	Experience (in years)
Machine Learning Architect	Hewlett Packard	17
AI Research Lead II	Wipro Tech	13
Associate Professor	Academician	23
Senior Data & Applied Scientist	Microsoft	12
Assistant Professor, Data Engineer	Academician	5
Senior Technology Consultant/ex-Research Scientist	Ernst & Young Global	5
Data analytics strategy	Ernst & Young Global	7
Assistant Professor	Academician	7
Assistant Professor	Academician	5

The evaluation score for the ith alternative is AS_i. The options are then sorted by considering the descending values of the appraisal scores.

Data Collection

The data for the present study have been collected at two different levels. This involves collecting opinions from chatbot experts on the criteria to evaluate the different alternatives through the Delphi method and information about the usefulness of the identified alternatives on the listed criteria from a second panel of experts. The data from the Delphi method were used to calculate the criteria weights through the CRITIC method. WASPAS methodologies were applied to rank chatbots on the alternatives and calculate the final overall rank of alternatives, and the EDAS technique validated the results. We discuss each step in more detail in this section.

Delphi was used for criteria identification (Delbecq et al., 1975). Delphi is a structured communication method used in research to gather expert opinions and reach a consensus among the experts (Delbecq et al., 1975). The Delphi method solves complex decision-making problems by systematically collecting, analyzing, and synthesizing expert opinion and literature data (Bouraima et al., 2023). It allows us to combine the combined list of criteria identified with the practical usability of the same by taking in expert opinions.

We identified a group of 10 experts, referred to as Expert Panel 1, with expertise in the chatbot industry and at least five years of experience. To encourage honest responses, we assured the participants of anonymity and confidentiality of their personal information. The details of the panel experts are presented in Table 3. We created a list of potential criteria and their category based on the classification discussed in the literature-review section. Each criterion was asked to be rated on a Likert scale of 1 to 5, where 1 represents the least importance, and 5 illustrates the highest importance for a criterion.

A list of questionnaires was compiled and distributed to a panel of experts. These included a list of all the categories of the evaluation matrix along with their questionnaire. The questionnaires were distributed via email, and the experts were asked to send back the filled survey. The expert opinion was then collected from the responses and analyzed. After compiling and analyzing data from the first round of reactions, a revised questionnaire was sent out to the panel of experts for another round of feedback. The revised responses were used to finalize the criteria list and calculate each criterion's relative importance via the CRITIC method.

A second round of data collection was performed to rank alternatives on each criterion. The target audience for this round study is experts from different industries with experience in using chatbots for task completion, referred to as Expert Panel 2.

Table 4. Expert Panel 2

Designation	Background	Experience (in years)
Credit Risk Analyst	JPMorgan Chase & Co.	3
Teaching Associate, Research scholar	Academician	5
Director Of Engineering	Trell	10
Software Developer	Inoweave	5
Research Scholar	Academician	3
Client Partner - Utilities	Tata Consultancy Services	20
Senior Consultant	FMC Technologies	9
Assistant Professor	Academician	5
Research Scholar	Academician	3

The data for this round have been collected by employing the survey questionnaire finalized in the previous round. Questions about each identified criterion were asked, and the respondents rated their experience with the platform on the specific factor mentioned in the question from 1 to 5, with 5 being the best and 1 being the worst. Similar to Round 1, the questionnaires and their responses were communicated over email. Nine experts were considered for the sample. These individuals were considered experts in using chatbots based on their experience with the use of chatbots. The average age of these respondents is 33.56. The details of the experts are mentioned in Table 4, with information about their backgrounds and years of experience in their respective industries.

RESULTS

The current study recommends a hybrid methodology with steps for evaluating the chatbot platform, as discussed in the previous sections. This model aims to determine the relative significance of many elements contributing to a platform's quality and then rate the chatbot platform according to its performance against those criteria. In this section, we discuss the results for the alternatives identified, followed by criteria and their weights and, finally, the ranking of the identified alternatives.

Alternatives

For alternative selection, we first searched for different AI-powered natural language processing tools available online. After selecting an initial list of alternatives, we listed each website's total number of visits from its inception. We collected information about the visits and ranked the alternatives. Figure 4 displays the visitors of each alternative. The visitors for alternatives other than ChatGPT and Character AI were not visible due to the massive difference in the number of visitors among these options. For comparison, we have provided a magnified version of the graph without these two options as part of Figure 4.

Next, we conducted a literature survey for the top alternatives to identify if they have been explored in extant research studies. A list of five alternatives was finalized after analysis of various alternative usage among users of chatbots, as shown in Table 5 (as of August, 2023). The table lists the details of these alternatives, with the website links, their launch date, the total number of visitors shown in Figure 4, and the references for the articles in the literature that have explored these alternatives. We discuss these alternatives in more detail in this section.

Table 5. Final List of Alternatives

Name	URL	Launched	Monthly Visits*	References
ChatGPT	https://chat.openai.com/	Nov 2022	1.4B	Van Dis et al., 2023; Zhu et al., 2023
Chatsonic	https://writesonic.com/chat	Dec 2022	7.6M	Chaka, 2023; Zhu et al., 2023
Perplexity AI	https://www.perplexity.ai/	Aug 2022	27.9M	Chaka, 2023; Zhu et al., 2023
Character AI	https://beta.character.ai/	Sep 2022	196.4M	Zou et al., 2023
You.com	https://you.com/	Nov 2021	13.4M	Chaka, 2023; Zhu et al., 2023

Note. Source: https://www.similarweb.com/

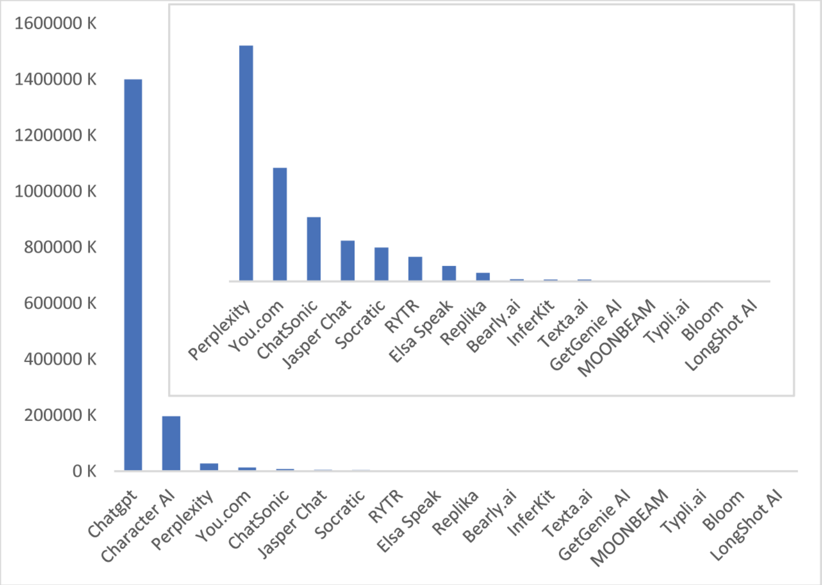
ChatGPT

ChatGPT is an AI-driven natural language processing tool that facilitates conversational engagements with chatbots across diverse scenarios. Users can respond to this language model and receive assistance composing emails, articles, and code (Van Dis et al., 2023; Zhu et al., 2023). It has garnered substantial attention in both contemporary media and academic discourse.

Chatsonic

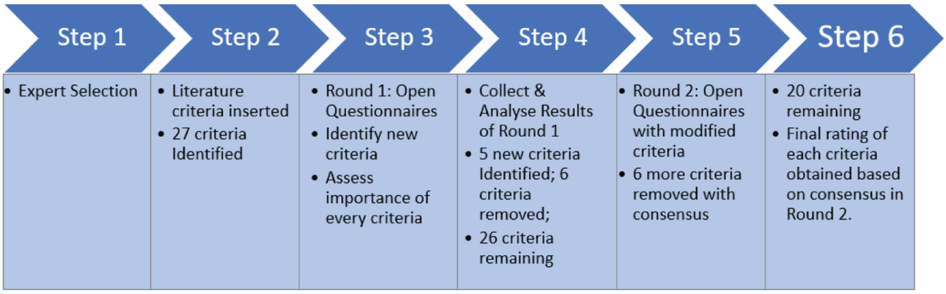
Chatsonic is an AI language model focusing on world affairs. While ChatGPT's database stopped in 2021, Chatsonic's assistance from Google means it can keep up with breaking news and deliver relevant responses and articles (Chaka, 2023; Zhu et al., 2023). Chatsonic can also generate AI-based images and integrate them with search engines to deliver real-time content. It provides extensive customization features and a user-friendly interface.

Figure 4. Chatbot Total Visitors



Note. Source: https://www.similarweb.com/

Figure 5. Schematic Flow Diagram of Delphi Research Framework


Step 1
Expert Selection
Step 2
Literature criteria inserted
27 criteria Identified
Step 3
Round 1: Open Questionnaires
Identify new criteria
Assess importance of every criteria
Step 4
Collect & Analyse Results of Round 1
5 new criteria Identified; 6 criteria removed;
26 criteria remaining
Step 5
Round 2: Open Questionnaires with modified criteria
6 more criteria removed with consensus
Step 6
20 criteria remaining
Final rating of each criteria obtained based on consensus in Round 2.

You

Google and other ad-supported search engines compete with You.com, a new search engine established by R. Socher, former Salesforce chief scientist. The search engine is created with sophisticated natural language processing to provide highly relevant results and business claims, and it will not depend on advertising for revenue. It is believed to include privacy settings, reliable user feedback, and thorough AI-powered analyses (Chaka, 2023; Zhu et al., 2023).

Perplexity

Perplexity AI originates from creators of an AI-powered search engine platform supporting big language models and results databases. The platform facilitates the direct construction of moderated and helpful AI language assistance and offers an open-source environment accessible to the public (Chaka, 2023; Zhu et al., 2023).

Character AI

Character AI is AI software that facilitates online conversations with nonhuman entities. Its AI system can take on the character of various professionals such as lawyers, doctors, and trainers (Zou et al., 2023). This AI system has garnered much user interest as it can create lifelike characters based on specific jobs or industries.

Identify Criteria

Building and adapting from the classification Russell-Rose (2017) gave, we grouped the chatbot traits observed in the selected studies into four main views or perspectives. The different categories of these perspectives have been carefully chosen to encompass the various aspects of human-chatbot interactions. The final questionnaire was developed and took the opinions of experts who have at least five years of experience using chatbots through the Delphi method, as explained in the methodology section. Figure 5 presents the framework for the Delphi Research, listing the number of criteria in each step of the process.

A final list of 20 criteria was selected for the study. All these criteria were identified by doing an extensive literature review and taking the experts' opinions. Table 6 lists the survey questions and the relevant details, such as the attribute measured, the perspective, and the sources from which the questionnaires were chosen for each criterion.

Relative Weights of the Factors

The next step is to determine the weights for each of the criteria. Combining all of the responses yielded a matrix of all the criteria. The responses were combined using a geometric mean approach.

Table 6. Final List of Criteria

Criteria	Description	Attribute	Perspective	References
C1	The chatbot's user interface was intuitive and easy to navigate	Navigability	User experience	(Eren, 2021)
C2	The chatbot was able to understand my questions and needs.	Ability to detect meaning and intent	Information retrieval	(Torres et al., 2019)
C3	The chatbot provided accurate and helpful information.	Adequateness of information	Information retrieval	(Chuah et al., 2021)
C4	The chatbot was able to provide appropriate feedback or follow-up.	Relevancy of responses to the context	Linguistic	(Chuah et al., 2021)
C5	The chatbot had a friendly and conversational tone.	Personality	User experience	Expert
C6	The chatbot provided the solutions to my problems.	Ability to aid	Information retrieval	(Mulyono et al., 2022)
C7	The chatbot was able to learn from my interactions and improve its responses.	Ability to learn	Technology	(Homes et al., 2019)
C8	The chatbot was able to provide relevant links or resources to assist me further.	Maintaining themed discussion	Information retrieval	Expert
C9	The chatbot was able to recognize and respond to humor or sarcasm appropriately.	Social relations	Linguistic	(Mulyono et al., 2022)
C10	The chatbot is integrated with other channels or platforms for better support.	Getting assistance	Technology	(Perez-Soler et al., 2021)
C11	The chatbot added value to my experience.	Human assistance	User experience	(Sidaoui et al., 2020)
C12	The chatbot response time was acceptable.	Response time frequency	Technology	(Shawar and Atwell, 2005)
C13	The chatbot handles my personal information securely	Privacy protection	User experience	(Abdulquadri et al., 2021)
C14	The chatbot provides me with quick and accurate answers to frequently asked questions.	Responding to specific questions	Information retrieval	(Mittal et al., 2021)
C15	I would recommend the chatbot to others based on my experience.	Impression	User experience	Expert
C16	It was easy to integrate the chatbot into my existing system.	Getting assistance	Technology	(Vegesna et al., 2018)
C17	The chatbot was able to provide clear instructions or guidance.	Understanding rate	Technology	(Følstad et al., 2018)
C18	The chatbot detected and corrected any errors or misunderstandings.	Unambiguity of the responses	Linguistic	(Chuah et al., 2021)
C19	The chatbot maintained the conversation even when I deviated from the main topic.	Maintaining themed discussion	Linguistic	(Perez, 2020)
C20	The chatbot was able to engage in natural and fluid conversation.	Naturalness	Linguistic	(Kirakowski et al., 2009)

The CRITIC technique is used to calculate the criteria weights under this section. The decision matrix is initially normalized. Table 7 displays the weight of each criterion. All of the criteria's standard deviations were calculated. The criteria weights are derived using Equation 1, as explained at the beginning of the methodology section.

Once the weights have been determined, the next step is to evaluate the rank of the alternatives based on the criteria.

Table 7. Weight of Criteria

Criteria	Weights	Criteria	Weights
C1	0.0435	C11	0.0534
C2	0.0499	C12	0.0338
C3	0.0435	C13	0.0738
C4	0.0384	C14	0.0694
C5	0.0598	C15	0.0524
C6	0.0499	C16	0.0452
C7	0.0599	C17	0.0427
C8	0.0570	C18	0.0397
C9	0.0323	C19	0.0586
C10	0.0497	C20	0.0470

Table 8. Evaluation Results for Each Chatbot

	WSM	WPM	Joint generalized criteria (lambda= 0.5)	Final Rank
[YOU]	3.157	1.936	2.547	4
[CHATGTP]	3.687	2.116	2.902	1
[PERPLEXITY]	3.3547	2.008	2.681	3
[CHATSONIC]	3.357	2.024	2.691	2
[CHARACTER AI]	2.935	1.865	2.400	5

Ranking the Alternatives

WASPAS

WASPAS is employed to determine the final rankings of chatbots based on their responses. The CRITIC weighted criterion is used with the WASPAS method to rank the alternatives. Zavadskas et al. (2012) introduced the WASPAS approach, utilizing weighted product and sum models. The first step of this method involves normalizing the decision matrix, as detailed in the second part of the methodology section, to ensure comparability. Subsequently, the weighted sum and product scores are computed for each alternative. For example, for the alternative "YOU," the scores are 3.120915 and 3.106297, respectively, while for ChatGPT, the scores are 3.617535 and 3.59174. Employing the WASPAS method allows us to rank these options and identify the optimal choice based on the provided criteria and their associated weights through aggregation functions.

A more general equation for assessing the relative significance of alternatives is utilized to enhance ranking accuracy and decision-making support in the WASPAS method. The evaluation results for each chatbot are shown in Table 8. WSM is calculated using Equation 1, as explained in the second part of the methodology section. Similarly, WPM is calculated using Equation 2, and the joint generalized criteria are calculated using Equation 4. ChatGPT was ranked the best overall platform, ranking 1 in 17 criteria out of 20.

EDAS

EDAS was used to validate the results obtained through WASPAS. The weights acquired using the CRITIC approach were applied to each criterion. The PDA and NDA were computed taking into account the type of criteria. The weighted sum product of PDA and NDA was then calculated for each possibility. The weighted ratings were normalized before the final appraisal score (ASi) was

Table 9. Evaluation by EDAS

	WPi	WNi	NWPi	NWNi	ASi	Final Rank
[You]	0.004	0.049	0.034	0.656	0.345	4
[CHATGTP]	0.113	0.002	1.000	0.989	0.994	1
[PERPLEXITY]	0.042	0.013	0.372	0.909	0.640	3
[CHATSONIC]	0.051	0.005	0.452	0.962	0.707	2
[CHARACTER AI]	0.000	0.142	0.000	0.000	0.000	5

Table 10. Overall Rank for Each Chabot

CHATBOTS	RANK (WASPAS)	RANK (EDAS)
[You]	4	4
[CHATGTP]	1	1
[PERPLEXITY]	3	3
[CHATSONIC]	2	2
[CHARACTER AI]	5	5

computed to rank the options. ChatGPT ranks highest after considering all the factors, making it the ideal option.

After data collection and normalization for each criterion, the average values for the criteria are computed, as explained in the third part of the methodology section. The values presented in Table 9 illustrate the relative rankings of each entity based on their performance. In this context, lower values in WPi (Equation 5), WNi (Equation 6), NWPi (Equation 7), and NWNi (Equation 8) imply superior performance, while a higher value in Asi (Equ ation 9) denotes a more favorable overall ranking. For example, ChatGPT exhibits the highest ASi value of 0.992867, signifying its status as the most proficient performer among the possibilities listed, as determined by the EDAS evaluation method.

Each chatbot's overall rank is calculated using the WASPAS first, and the EDAS technique is used subsequently for validation, as shown in Table 10.

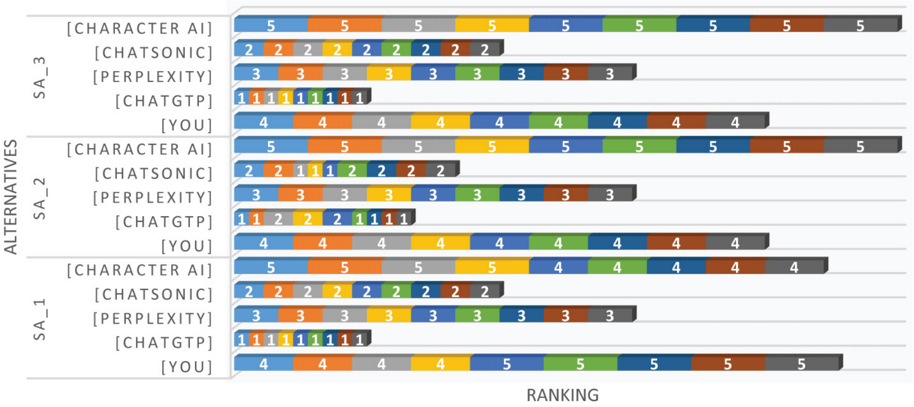
Based on the overall scores, ChatGPT is the best alternative among the five chatbots evaluated. Chatsonic and Perplexity are also ranked highly, while YOU and Character AI are ranked lower. The results were sent to Panel 2 experts for final comments. All the experts agreed with the definitive ranking of chatbots.

SENSITIVITY ANALYSIS

When applying MCDM, the results' validity, robustness, and generalizability are improved with sensitivity analysis. It increases transparency, offers a more thorough grasp of the decision issue, and promotes the development of the subject's body of knowledge. Hence, the sensitivity analysis is included to enhance the quality of the results of this paper.

The output from the previous section allows us to determine the priority of the alternatives based on the assigned weights of the criteria and alternatives. The robustness of the result is studied by performing a sensitivity analysis of the model, taking into account the uncertainties in the input (Saltelli, 2002). To perform sensitivity analysis, the relative weights of the criteria are modified, and the weights of the criteria for the alternatives are modified, or both are modified simultaneously. By doing so, one can examine how these changes impact the priorities assigned to different alternatives. Sensitivity analysis focuses on understanding the impact of adjusting the relative criteria weights on the ranking. This section examines how these changes affect the overall ranking.

Figure 6. Results of Sensitivity Analysis for WASPAS



The stability of the outcomes is examined using a sensitivity analysis over a wide range of input variable values. Twenty subfactors are included in the present research, and analyzing 20 different weight patterns for sensitivity analysis becomes impractical. In such a scenario, the top 10 subfactors can be chosen to see the impact of their changes on the overall results for the sensitive analysis (Sharma et al., 2020).

For sensitivity analysis, one can generate statistical estimates of rankings by repeatedly simulating estimations of the decision matrix elements within a given margin of error for computing the ranks of options. The analysis was conducted for different lambda values starting from 0.10 to 1. The rankings were consistent across all values, proving the platform's ranking method's robustness.

The stability of the findings was examined by putting the model to the test using ten alternative weights for the top 10 subcriteria (designated by SA1-SA10). As shown in Table 11, one subfactor has the lowest weight, and one subfactor has the most significant weight in a given combination of weights. According to the sensitivity analysis findings, the alternatives' rankings hold steady across nine of the 10 sets of weights.

As shown in Table 11, one component has the lowest weight, and one factor has the highest weight in a specific set of weights. According to the sensitivity analysis findings, the alternatives' rankings hold for all weights (Figure 6). The top alternatives have continuously been ChatGTP and Chantsonic, while YOU and Character AI have constantly ranked among the worst. We also created a sensitivity analysis test depending on the value to confirm the outcomes of the WASPAS approach. Using various values, we solved the decision-making model 10 times. The results were consistent with Table 10, and the platform's rankings remained unchanged.

DISCUSSION AND IMPLICATIONS

With the recent technological advancements in learning language models, chatbots have been widely adopted for task completion. The current study investigates the importance of chatbot features, which play a role in its adoption. We start with listing essential criteria for the evaluation of chatbots and usability. Four broad categories of criteria were adopted from existing literature. These are user experience, information retrieval efficiency, linguistic capabilities, and technological perspective. User experience was the most crucial category, followed by linguistic capabilities, technical perspective, and information retrieval efficiency.

This reflects the focus on creating user-centric chatbots that effectively meet user needs. User experience is paramount because chatbots are designed to interact with users helpfully and pleasantly. A chatbot with a poor user experience can frustrate users, leading to abandonment or

Table 11. Ranking Order With Different λ Value

λ	SA_1					SA_2					SA_3				
	[You]	[ChatGPT]	[Perplexity]	[Chatsonic]	[Character AI]	[You]	[ChatGPT]	[Perplexity]	[Chatsonic]	[Character AI]	[You]	[ChatGPT]	[Perplexity]	[Chatsonic]	[Character AI]
0.1	4	1	3	2	5	4	1	3	2	5	4	1	3	2	5
0.2	4	1	3	2	5	4	1	3	2	5	4	1	3	2	5
0.3	4	1	3	2	5	4	2	3	1	5	4	1	3	2	5
0.4	4	1	3	2	5	4	2	3	1	5	4	1	3	2	5
0.5	5	1	3	2	4	4	2	3	1	5	4	1	3	2	5
0.6	5	1	3	2	4	4	1	3	2	5	4	1	3	2	5
0.7	5	1	3	2	4	4	1	3	2	5	4	1	3	2	5
0.8	5	1	3	2	4	4	1	3	2	5	4	1	3	2	5
0.9	5	1	3	2	4	4	1	3	2	5	4	1	3	2	5
1.0	5	1	3	2	4	3	2	4	1	5	4	1	3	2	5

dissatisfaction. The chatbot should be able to understand and interpret user input accurately, including recognizing slang, colloquialisms, and various accents. It should engage users in coherent, context-aware conversation, ensuring smooth and intuitive interaction. The chatbot should be able to tailor responses and recommendations based on user preferences and historical interactions, creating a more personalized experience. An effective chatbot should also have some emotional intelligence, recognizing and responding to user emotions appropriately.

After a good user experience, the chatbot's linguistic capabilities are crucial. This includes the chatbot's ability to use language effectively and persuasively. The chatbot should generate grammatically correct sentences and use proper language structure. It should have a vast vocabulary and use appropriate language based on the context and user. The chatbot's tone and style should align with its brand or purpose, and it should be able to switch between formal and informal language as needed.

The technological perspective highlights the importance of the underlying technology that powers the chatbot. A chatbot must be built on a solid technical foundation to ensure reliability and scalability. The chatbot should be able to handle a growing user base and increase conversation volume without performance issues. Protecting user data and ensuring secure communication is crucial. The chatbot should seamlessly integrate with other systems, platforms, or databases to provide relevant information or services.

While user experience and linguistic capabilities are user-facing aspects, information retrieval efficiency ensures that the chatbot can quickly and accurately fetch the information or perform tasks that users request. The chatbot should respond promptly, minimizing user wait times. It should retrieve and present information accurately, reducing errors and misunderstandings.

By prioritizing these criteria, chatbots will excel in their core functions and provide a pleasant and effective user experience. However, it is important to note that the specific priorities may vary depending on the chatbot's intended use case and target audience. The participants in this study were working adults and professionals. According to the results, multiple criteria are essential for determining the user preferences for a language-related task completion chatbot. Customers want better and quicker service and an easier way to work, which is more time-efficient.

This research paper evaluated five different chatbot models based on the identified criteria. It was determined that ChatGPT was excellent at assisting users with their questions. It ranked first in 17 out of 20 criteria, demonstrating its wide usability across different domains. ChatSonic was ranked first in 2 of 20 criteria. This included the conversational tone and the speed of the chatbot. The ranking of other factors was a mix of different positions with the worst performance in its ability to learn. This indicates that, unlike ChatGPT, ChatSonic is suitable for more superficial interactions, which are closer to standard FAQs and excellent at generating quick answers. Perplexity AI was ranked second in most criteria overall. It ranked first in providing citations and links to the sources for its answers. This indicated its usability in applications where the source of information also becomes necessary, such as research and education. Character AI was last overall and performed worst in most criteria but tied with ChatGPT in the first position in privacy perceptions. Finally, YouChat was fourth overall and had a mix of positions with no criteria where it ranked first. These results stress the significance of tailoring the chatbot model choice to the needs of each application.

Businesses and organizations increasingly use chatbots to communicate with and support their clientele. Chatbots are getting smarter and more intelligent and can now understand and respond to natural language queries. Chatbots are becoming sophisticated and effective tools for customer engagement and support. While chatbots cannot fully replace human interaction, they can help speed up and simplify mundane tasks so that human agents can focus on higher-value, more complex tasks.

Managerial Implications

This research provides new insights for both academics and practice. Chatbots have already made their debut and will remain a fixture in the industry, but it is unclear how their quality should be evaluated. Our research adds to the theoretical literature by demonstrating that the quality of responses from chatbots is a crucial factor. It confirms that multiple factors form essential attributes of successful chatbots and should be included as one of the observable characteristics in chatbot evaluations. Another theoretical addition is the analysis and exploration of specific criteria based on classification adopted from the literature. While ChatGPT is ranked one overall, it performs lower on some criteria such as speed, providing relevant sources, and integration into other platforms. This demonstrated the importance of including multiple criteria while evaluating chatbot importance.

Theoretical Implications

This study's most important addition to management is identifying the factors that contribute to the positive user experience of chatbots. Designers of chatbot architecture and managers will find this knowledge beneficial since they are tasked with encouraging these characteristics in destination chatbots. Regarding the use of chatbot alternatives, while ChatGPT performs better than other chatbots overall, individual differences in the criteria ranking have practical implications for their use. For instance, Perplexity might become more relevant in educational and academic contexts where the sources are important to provide support to the claims and add credibility to the responses. On the other hand, ChatGPT is more relevant for customer-facing chatbots where user experience is paramount. Similarly, ChatSonic becomes more suitable for scenarios where the speed of answer generation is important.

CONCLUSION

While this study provides some interesting findings, it suffers from certain limitations. The small sample size and narrow focus suggest it may be a pilot for a more extensive, comprehensive study. This study's supplementary data suggest further investigation into various factors, such as the medical illnesses being studied, the specialists in many sectors interacting with chatbots, and cultural norms and practices. One possibility is to look into the viewpoints of professionals in specific industries and see whether they vary from those of professionals in other sectors and services that use chatbots. Future research may show how individuals working on similar initiatives are disconnected from one another and highlight the need for more dialogue.

MCDM was used in this study to rank the alternatives, but it has several limitations. First, the MCDM technique assumes that the criteria used are independent, which may not always be accurate in real-world scenarios. Second, assigning weights to each criterion is subjective and may vary depending on the decision-maker's preferences, which may introduce some bias into the evaluation. Finally, the MCDM technique may not be suitable for situations where the criteria are subjective and difficult to quantify.

Despite these limitations, the MCDM technique provides a valuable framework for decision-making in situations with multiple criteria to evaluate. The MCDM technique in this study helped objectively determine each criterion's relative weights. The paper identifies essential criteria for chatbot evaluation using the Delphi method and evaluates the relative importance of each criterion using the CRITIC method. Furthermore, we identify widely used alternatives based on the number of visitors for the top online chatbots and rank them using the WAPAS and EDAS methods. The current study lays a strong foundation for future exploration and expansion. There is an exciting opportunity to incorporate additional open AI alternatives, ensuring robustness through meticulous consideration of utility. Finally, integrating a fuzzy MCDM model promises to elevate the robustness of the process even further. This forward-looking approach holds immense potential for advancing the scope and impact of our research.

ACKNOWLEDGMENT

The study was supported by:

1. Zhejiang Social Science Planning Project: “Research on the Ecological Index and Collaborative Mechanism of Ecological Governance of Port and Shipping Logistics on the Maritime Silk Road” (19NDJC280YB)
2. Ministry of Education of China (MOE) Project of Humanities and Social Science Foundation: “Research on the Evolutionary Game Mechanism and Collaborative Mechanism of Cluster Green Supply Chain: A Case Study of Typical Industries in Zhejiang Province” (19YJA790041)
3. Key Research Institute of Philosophy and Social Sciences of Zhejiang Province–Modern Port Service Industry and Creative Culture Research Center of Zhejiang Wanli University

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING STATEMENT

No funding was received for this work.

PROCESS DATES

Received: November 6, 2023, Revision: March 19, 2024, Accepted: April 9, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Qiuzheng Li; lqz@zwu.edu.cn

REFERENCES

Abdulquadri, A., Mogaji, E., Kieu, T. A., & Nguyen, N. P. (2021). Digital transformation in financial services provision: A Nigerian perspective to adopting chatbot. *Journal of Enterprising Communities: People and Places in the Global Economy*, *15*(2), 258–281. 10.1108/JEC-06-2020-0126

AbuShawar, B., & Atwell, E. (2016). Usefulness, localizability, humanness, and language-benefit: Additional evaluation criteria for natural language dialogue systems. *International Journal of Speech Technology*, *19*(2), 373–383. 10.1007/s10772-015-9330-4

Bitner, M. J., Ostrom, A. L., & Meuter, M. L. (2002). Implementing successful self-service technologies. *The Academy of Management Perspectives*, *16*(4), 96–108. 10.5465/ame.2002.8951333

Bouraima, M. B., Tengecha, N. A., Stević, Ž., Simić, V., & Qiu, Y. (2023). An integrated fuzzy MCDM model for prioritizing strategies for successful implementation and operation of the bus rapid transit system. *Annals of Operations Research*, ●●●, 1–32. 10.1007/s10479-023-05183-y36743351

Cai, W., Jin, Y., & Chen, L. (2022). Task-oriented user evaluation on critiquing-based recommendation chatbots. *IEEE Transactions on Human-Machine Systems*, *52*(3), 354–366. 10.1109/THMS.2021.3131674

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*, *6*(2).

Chakrabortty, R. K., Abdel-Basset, M., & Ali, A. M. (2023). A multi-criteria decision analysis model for selecting an optimum customer service chatbot under uncertainty. *Decision Analytics Journal*, *6*, 100168. 10.1016/j.dajour.2023.100168

Chuah, K. M., & Kabilan, M. (2021). Teachers' views on the use of chatbots to support English language teaching in a mobile environment. [iJET]. *International Journal of Emerging Technologies in Learning*, *16*(20), 223–237. 10.3991/ijet.v16i20.24917

Chwelos, P., Benbasat, I., & Dexter, A. S. (2001). Empirical test of an EDI adoption model. *Information Systems Research*, *12*(3), 304–321. 10.1287/isre.12.3.304.9708

Dale, R. (2016). The return of the chatbots. In *Natural language engineering* (Vol. 22, No. 5, pp. 811–817). 10.1017/S1351324916000243

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning: A guide to nominal group and Delphi processes*. Scott, Foresman.

Diakoulaki, D., Mavrotas, G., & Papayannakis, L. (1995). Determining objective weights in multiple criteria problems: The critic method. *Computers & Operations Research*, *22*(7), 763–770. 10.1016/0305-0548(94)00059-H

Eren, B. A. (2021). Determinants of customer satisfaction in chatbot use: Evidence from a banking application in Turkey. *International Journal of Bank Marketing*, *39*(2), 294–311. 10.1108/IJBM-02-2020-0056

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In *Internet science: 5th international conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, proceedings 5* (pp. 194–208). Springer International Publishing. 10.1007/978-3-030-01437-7_16

Hsu, M. C. (2023). The construction of critical factors for successfully introducing chatbots into mental health services in the army: Using a hybrid MCDM approach. *Sustainability (Basel)*, *15*(10), 7905. 10.3390/su15107905

Huang, L. P., Sing, G. O., Kumar, Y. J., & Pradana, A. (2018). A black-box approach evaluation on conversational agent using Loebner prize competition datasets.

Jati, H., & Wardani, R. (2021). Visibility ranking of university e-learning websites: CRITIC method approach. []. IOP Publishing.]. *Journal of Physics: Conference Series*, *1737*(1), 012030. 10.1088/1742-6596/1737/1/012030

Keeney, R. L. (1992). *Value-focused thinking: A path to creative decision-making*. Harvard University Press.

Kirakowski, J., O'Donnell, P., & Yiu, A. (2009). Establishing the hallmarks of a convincing chatbot-human dialogue. *Human-Computer Interaction*, ●●●, 49–56.

Mahesh, B. (2020). Machine learning algorithms: A review. [IJSR]. *International Journal of Scientific Research*, *9*(1), 381–386.

Meisel, W. (n.d.). Building digital assistants and bots: A vendor guide and market analysis. In *Tech. Rep*. (p. 400). https://www.aitrends.com/research-bots/

Mittal, M., Battineni, G., Singh, D., Nagarwal, T., & Yadav, P. (2021). Web-based chatbot for frequently asked queries (FAQ) in hospitals. *Journal of Taibah University Medical Sciences*, *16*(5), 740–746. 10.1016/j.jtumed.2021.06.00234690656

Mokmin, N. A. M., & Ibrahim, N. A. (2021). The evaluation of chatbot as a tool for health literacy education among undergraduate students. *Education and Information Technologies*, *26*(5), 6033–6049. 10.1007/s10639-021-10542-y34054328

Mulyono, J. A., & Sfenrianto, S. (2022). Evaluation of customer satisfaction on Indonesian banking chatbot services during the COVID-19 pandemic. *CommIT (Communication and Information Technology). Journal*, *16*(1), 69–85.

Narducci, F., Basile, P., de Gemmis, M., Lops, P., & Semeraro, G. (2020). Investigating the user interaction modes of conversational recommender systems for the music domain. *User Modeling and User-Adapted Interaction*, *30*(2), 251–284. 10.1007/s11257-019-09250-7

Orden-Mejía, M. A., & Huertas, A. (2022). Tourist interaction and satisfaction with the chatbot evokes pre-visit destination image formation? A case study. *Anatolia*, 1–15. Gulum, P., Ayyildiz, E., & Gumus, A. T. (2021). A two level interval valued neutrosophic AHP integrated TOPSIS methodology for post-earthquake fire risk assessment: An application for Istanbul. *International Journal of Disaster Risk Reduction*, *61*, 102330.

Peng, G., & Bhaskar, R. (2023). Artificial intelligence and machine learning for job automation: A review and integration. [JDM]. *Journal of Database Management*, *34*(1), 1–12. 10.4018/JDM.318455

Perez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549–1565. 10.1002/cae.22326

Perez-Soler, S., Juarez-Puerta, S., Guerra, E., & de Lara, J. (2021). Choosing a chatbot development tool. *IEEE Software*, *38*(4), 94–103. 10.1109/MS.2020.3030198

Phooriyaphan, S., & Rachsiriwatcharabul, N. (2022). Identifying the criteria for selection of healthcare chatbot in Thailand using a multi-criteria decision making approach. *Journal of Positive School Psychology*, ●●●, 3757–3763.

Reiswich, A., & Haag, M. (2019). Evaluation of chatbot prototypes for taking the virtual patient's history. *dHealth*, 73–80.

Rogers, E. M. (1995). *Diffusion of innovations.* Free Press, *12*.

Rokhsaritalemi, S., Sadeghi-Niaraki, A., Kang, H. S., Lee, J. W., & Choi, S. M. (2022). Ubiquitous tourist system based on multi-criteria decision making and augmented reality. *Applied Sciences (Basel, Switzerland)*, *12*(10), 5241. 10.3390/app12105241

Russell-Rose, T. (2017, January 24). A framework for chatbot evaluation. Retrieved September 6, 2023, from https://isquared.wordpress.com/2017/01/24/a-framework-for-chatbot-evaluation/

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Analysis*, *22*(3), 579–590. 10.1111/0272-4332.0004012088235

Santa Barletta, V., Caivano, D., Colizzi, L., Dimauro, G., & Piattini, M. (2023). Clinical-chatbot AHP evaluation based on “quality in use” of ISO/IEC 25010. *International Journal of Medical Informatics*, *170*, 104951. 10.1016/j.ijmedinf.2022.10495136525800

Sharma, D., Srivastava, P. R., Pandey, P., & Kaur, I. (2020). Evaluating quality of matrimonial websites: Balancing emotions with economics. *American Business Review*, *23*(2), 9. 10.37625/abr.23.2.358-392

Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, *10*(4), 489–516. 10.1075/ijcl.10.4.06sha

Sidaoui, K., Jaakkola, M., & Burton, J. (2020). AI feel you: Customer experience assessment via chatbot interviews. *Journal of Service Management*, *31*(4), 745–766. 10.1108/JOSM-11-2019-0341

Singh, N., Chakraborty, A., Banik, B., Biswas, S. D., & Majumdar, M. (2022). Digital banking chatbots related MCDM problem by TODIM strategy in pentagonal neutrosophic arena. *Journal of Neutrosophic and Fuzzy Systems*, *4*(2), 26–41. 10.54216/JNFS.040202

Sugisaki, K., & Bleiker, A. (2020). Usability guidelines and evaluation criteria for conversational user interfaces: A heuristic and linguistic approach. In *Proceedings of mensch und computer 2020* (pp. 309–319). 10.1145/3404983.3405505

Syamsuddin, I., & Warastuti, S. W. (2021, December). Selecting chatbot platform for health enterprise training: A fuzzy AHP approach. In 2021 *international conference on decision aid sciences and application (DASA)* (pp. 756–760). IEEE.

Torres, C., Franklin, W., & Martins, L. (2019). Accessibility in chatbots: The state of the art in favor of users with visual impairment. In *Advances in usability, user experience and assistive technology: Proceedings of the AHFE 2018 international conferences on usability & user experience and human factors and assistive technology*, held on July 21–25, 2018, in Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9 (pp. 623–635). Springer International Publishing.

Triantaphyllou, E., & Mann, S. H. (1989). An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox. *Decision Support Systems*, *5*(3), 303–312. 10.1016/0167-9236(89)90037-7

Tung, A. (2021, November). A value focused thinking approach to decommissioning decision making. In *SPE Symposium: Decommissioning and Abandonment (p. D031S005R004)*. SPE. 10.2118/208471-MS

Turing, A. M. (2009). *Computing machinery and intelligence*. Springer Netherlands. 10.1007/978-1-4020-6710-5_3

Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. 10.1038/d41586-023-00288-736737653

Vegesna, A., Jain, P., & Porwal, D. (2018). Ontology based chatbot (for e-commerce website). *International Journal of Computer Applications*, *179*(14), 51–55. 10.5120/ijca2018916215

Yalcin, A. S., Kilic, H. S., & Delen, D. (2022). The use of multi-criteria decision-making methods in business analytics: A comprehensive literature review. *Technological Forecasting and Social Change*, *174*, 121193. 10.1016/j.techfore.2021.121193

Yel, İ., Sarucan, A., & Baysal, M. E. (2021, August). An application of fuzzy AHP, EDAS and WASPAS for the selection of process method in software projects. In *Intelligent and fuzzy techniques for emerging conditions and digital transformation:Proceedings of the INFUS 2021 conference, heldAugust 24-26, 2021.* Volume 1 (pp. 351–359). Springer International Publishing.

Yucenur, G. N., & Ipekçi, A. (2021). SWARA/WASPAS methods for a marine current energy plant location selection problem. *Renewable Energy*, *163*, 1287–1298. 10.1016/j.renene.2020.08.131

Zavadskas, E. K., Turskis, Z., Antucheviciene, J., & Zakarevicius, A. (2012). Optimization of weighted aggregated sum product assessment. *Elektronika ir Elektrotechnika*, *122*(6), 3–6. 10.5755/j01.eee.122.6.1810

Zhu, L., Mou, W., & Chen, R. (2023). Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *Journal of Translational Medicine*, *21*(1), 1–4. 10.1186/s12967-023-04123-537076876

Zou, S., Xu, Z., & Han, X. (2023). Research on embodiment and social robotics from the perspective of metacosmos: Research based on Character AI. In *SHS web of conferences* (*Vol. 168*). EDP Sciences. 10.1051/shsconf/202316803020

Harshit Kumar Singh is an Assistant Professor at the Indian Institute of Management Rohtak. He holds a PhD from the Indian Institute of Management Ahmedabad in Information Systems. Prior to joining his doctoral program, he worked in the industry as a systems developer. He had completed his Batcher's in Computer Science & Engineering. His research interest includes data structure and algorithms, IS competencies and motivational information systems focusing on gamification and digital engagement.

Surabhi Sakshi is a doctoral student at the Indian Institute of Management, Rohtak.

Justin Zhang is a faculty member in the Department of Management at Coggin College of Business in University of North Florida. He received his Ph.D. in Business Administration with a concentration on Management Science and Information Systems from Pennsylvania State University, University Park. His research interests include economics of information systems, knowledge management, electronic business, business process management, information security, and social networking. He has published research articles in various scholarly journals, books, and conference proceedings. He is the editor-in-chief of the Journal of Global Information Management. He also serves as an associate editor and an editorial board member for several other journals.

Qiuzheng Li, Professor, Vice Dean of the School of Logistics and E-commerce. PhD in Regional Economics from Nankai University, Postdoctoral Fellow at the Institute of Finance and Strategy, Chinese Academy of Social Sciences, Visiting Scholar at Darmstadt University of Technology in Germany. Research Area: Ecological governance of port and shipping logistics, Green supply chain management.