SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I4P109

Volume 12 Issue 4, 132-142, April 2025
© 2025 Seventh Sense Research Group®

Review Article

Trustworthiness Metrics for Measuring Efficiency of Chatbots - A Systematic Review

P. Sowmya[1], Vasudeva[2], Manjula Gururaj Rao[3]

[1,2,3]*Department of CSE, NMAM Institute of Technology, Nitte (Deemed to be University), Karkala, Udupi, Karnataka, India.*

[1]*Corresponding Author : sowmyap217@gmail.com*

Received: 06 February 2025 Revised: 08 March 2025 Accepted: 09 April 2025 Published: 29 April 2025

Abstract - *A chatbot acts as an AI-based virtual assistant for many applications like websites, banking apps, customer support systems and many more. It uses Artificial Intelligence (AI) to respond to users' queries without human intervention. In an application where there could be hundreds of options, searching for a specific option becomes a hassle for the user. Chatbot could solve all such problems, where chat with the bot and work done. However, when so many AI applications exist, it becomes critical to determine whether an AI application or tool is trustworthy. This article focuses on different evaluation metrics, ethical concerns and trustworthiness of AI applications, which help predict the efficiency of different AI-based chatbot systems.*

Keywords - *Artificial Intelligence, Chatbot, Evaluation, Natural Language Processing, Trust.*

1. Introduction

The imitation of human intelligence into machines, which are intended to think and comprehend just like humans, is popularly known as artificial intelligence, or AI. Artificial Intelligence (AI) provides various technologies and methodologies that allow computers and other machines to execute tasks that otherwise conventionally require human intelligence. Solving a Problem, learning with experience, identifying patterns, grasping natural language, and decision-making are a few of these tasks. [1]. Natural Language Processing (NLP), robotics, computer vision, deep learning, and machine learning are examples of AI technologies and methodologies.

Artificial Intelligence (AI) has many practical uses in a wide range of industries, including healthcare (diagnosis and treatment), finance (fraud detection and trading), driverless cars, chatbots for customer service, manufacturing (robotic automation), and many more. Along with the many benefits of artificial intelligence, numerous failures of AI have been documented throughout history. Some of these failures include misidentifications in facial recognition, incorrect recommendations in medical therapy, bias in decision-making, and loss in financial investments.

All of these failures could have a negative impact on human health and well-being. In order to prevent such enormous losses, finding an AI system's efficiency should be the top priority before using it. There are enormous uses for Artificial Intelligence (AI), but this article concentrates on one particular use: the virtual assistant, or chatbot, as it is more commonly called. A chatbot is a software or program imbibed in an application intended to communicate with users orally or in writing using natural language [2]. They can be found on multiple platforms, which include messaging apps, websites, and customer support systems. When relying on customer service for answers to our questions, drawbacks include long wait times before speaking with an executive agent, the service not being available around the clock, and the high cost of human-oriented customer service being high. One customer support representative can only speak with one customer at a time. Rather, the chatbot is available 24*7, and quick responses can be expected, multiple users can be responded simultaneously. Along with all other factors, the efficiency of the chatbot, that is, responses given to user's queries, plays a major role. Either the response should be right or almost right.

The right response can be achieved by choosing the best algorithms and evaluation metrics with respect to the context in which it is used. Currently, the chatbot evaluation is still dependent on human expert evaluation, which could be biased. The paper briefs out the algorithms and performance metrics that can contribute towards automatic evaluation and trustworthiness of chatbots.

2. Research Method

Figure 1 describes the flow of the contents of this article. This section gives a detailed overview of the different aspects of chatbots that make them trustworthy systems. Starting with composing research questions, research questions can be further categorized into identifying studies, analysis and also identifying report findings.

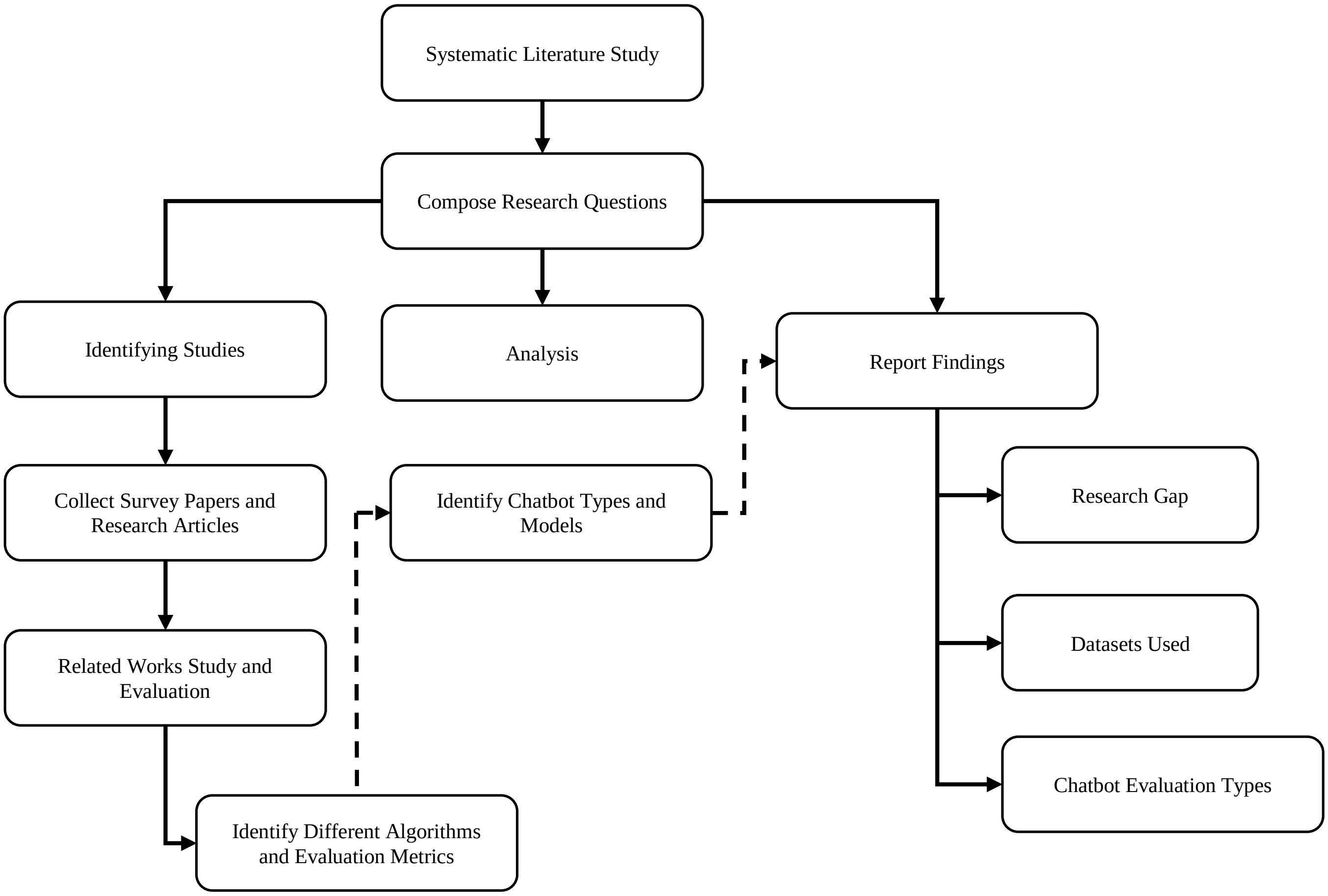
Systematic Literature Study
Compose Research Questions
Identifying Studies
Analysis
Report Findings
Collect Survey Papers and Research Articles
Identify Chatbot Types and Models
Research Gap
Related Works Study and Evaluation
Datasets Used
Chatbot Evaluation Types
Identify Different Algorithms and Evaluation Metrics

Fig. 1 Organization of systematic literature study

Identifying studies focuses on collecting articles based on Q1, indicating adequate papers that can be studied in this domain. Related works study, and evaluation focus on Q2, where, apart from survey papers, many studies were made on articles involving different algorithms, evaluation kinds, and techniques, and the consideration point is a chatbot.

In the analysis phase, based on all the studies from previous steps, an analysis was made about possible chatbot and model types based on Q3 and Q4. In the last phase, findings will be reported based on all the studies and analyses.

Findings were collected and reported, including datasets used, performance assessment based on Q5 and Q6, evaluation metrics based on Q7 and Q8, and different algorithm inspection techniques.

2.1. Research Question

When artificial intelligence was first being developed, scientists were primarily concerned with teaching computers to think like humans. However, this was no simple task. The human mind processes information by combining several aspects before deciding to carry out or act upon our views. Here, the variables affecting the selection are domain-specific and change depending on the domain, resulting in the creation of user-response applications, which prompted researchers to investigate potential avenues for developing chatbots and determine how best to generate responses using emerging technologies and diverse algorithms and methodologies. The authors looked at the topic's most recent study trends as well as the benefits and drawbacks of earlier studies. Then, as indicated in Table 1, two questions were framed to contribute towards data collection and the analysis criteria.

Table 1. Specific research questions aiming to analyze growth of chatbots in different domain

Question No	Question	Criteria of Evaluation
Q1	What is the research growth identified in the chatbot domain?	Total count of research papers from 2006 to 2023 on chatbot Figure 2.
Q2	What is current research knowledge in this domain?	Data analysis is done by applying different classifications, algorithms, evaluation techniques used, research gaps, and future directions.

2.2. Identifying Studies

Numerous research articles obtained for this paper's study are survey papers. At the same time, they are comparable to this research paper in nature because diverse technologies are utilized in each of these articles; they cannot all be classified into different tiers.

One study that was looked into includes a Survey on Chatbot Design Techniques in Speech Conversation Systems [8]. One research paper presented a Survey on Chatbot Evaluation Methods [17].

A survey on various algorithms used in chatbots examined a study on algorithms used in chatbot development [19]. A Literature Review of Recent Advances in Chatbots provides up-to-date information on chatbot advancements [26]. Figure 2 displays the distribution of publications in Scopus with the keywords "chatbot" or "chatbots" from 2006 to 2023 in ascending order. The number of research articles published on chatbots has dramatically increased year over year; in 2023, there were 863 research publications published on chatbots, up from 620 in 2022.

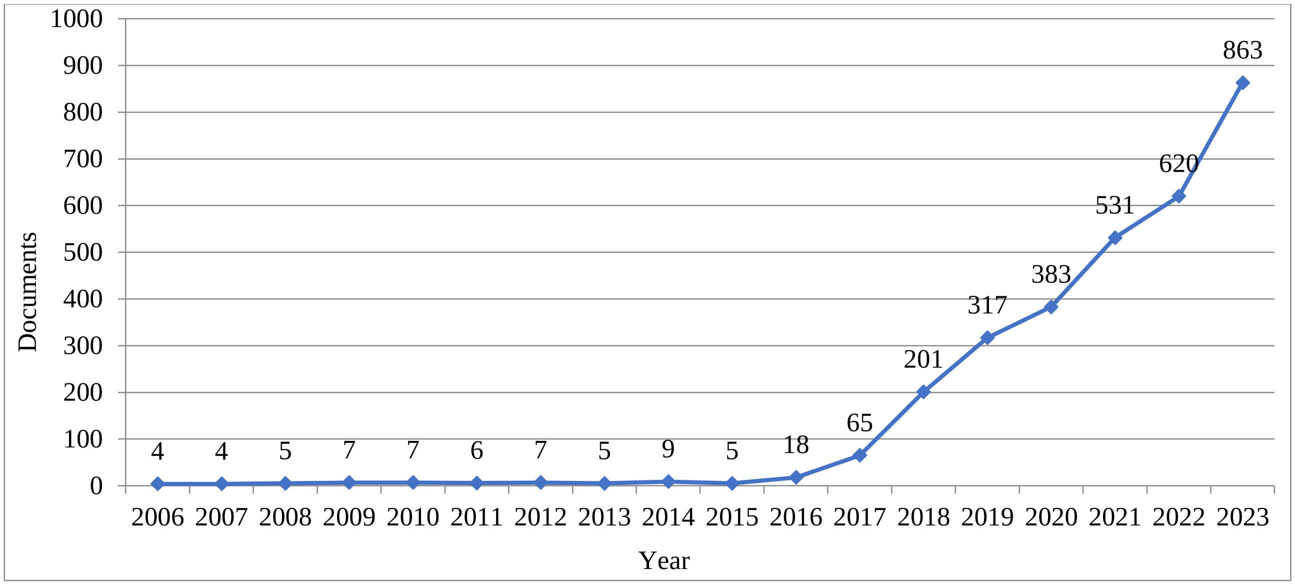

Fig. 2 Results of search, from 2006 to 2023 in Scopus with keywords “chatbot” or “chatbots” for Q1

2.3. Related Works Study and Evaluation

At this point, the writers reviewed each article's text to find any relevant studies being conducted on the subject. The articles that have been chosen for this paper discuss chatbot types, the requirement for evaluation using a variety of algorithms, and the necessary effective metrics. Chatbot Evaluation kinds from different articles are also considered, and it is appropriate to emphasize the type's efficiency. Every chosen article has a significant and original point that advances the related subject.

2.4. Analysis

2.4.1. Chatbot Types

A chatbot is a tried-and-true simulation of human thought processes, but it is not a human brain that can handle several tasks at once. Creating a chatbot with a certain goal in mind for a wide range of topics becomes necessary. In general, chatbots can be divided into two categories: domain-specific and social. Domain-specific chatbots are created with a specific purpose in mind, and their prompt and accurate responses are highly valued. On the other hand, social chatbots may be created specifically for interactions in which responses may be delayed.

Table 2. Defined research question on chatbot types based on research question in Table 1

Question No	Question	Criteria of Evaluation
Q3	How can the chatbots be divided based on their applications?	Latest chatbots of 2024 Table 3.
Q4	How can the chatbots be divided based on the model used?	Processing of input and generation of response Figure 3.

Additionally, the rule-based and generative models can be applied to chatbots based on how they process data and generate responses. Using a rule-based approach, they choose responses of the system by determining the input text lexical form and selecting it based on some preset group of predefined rules, all without creating any new text responses. Conversational rules are followed in the human hand-coding, structuring, and presentation of the knowledge used by the chatbot [28]. With a broad collection of rule libraries, chatbots can react to a greater range of user input.

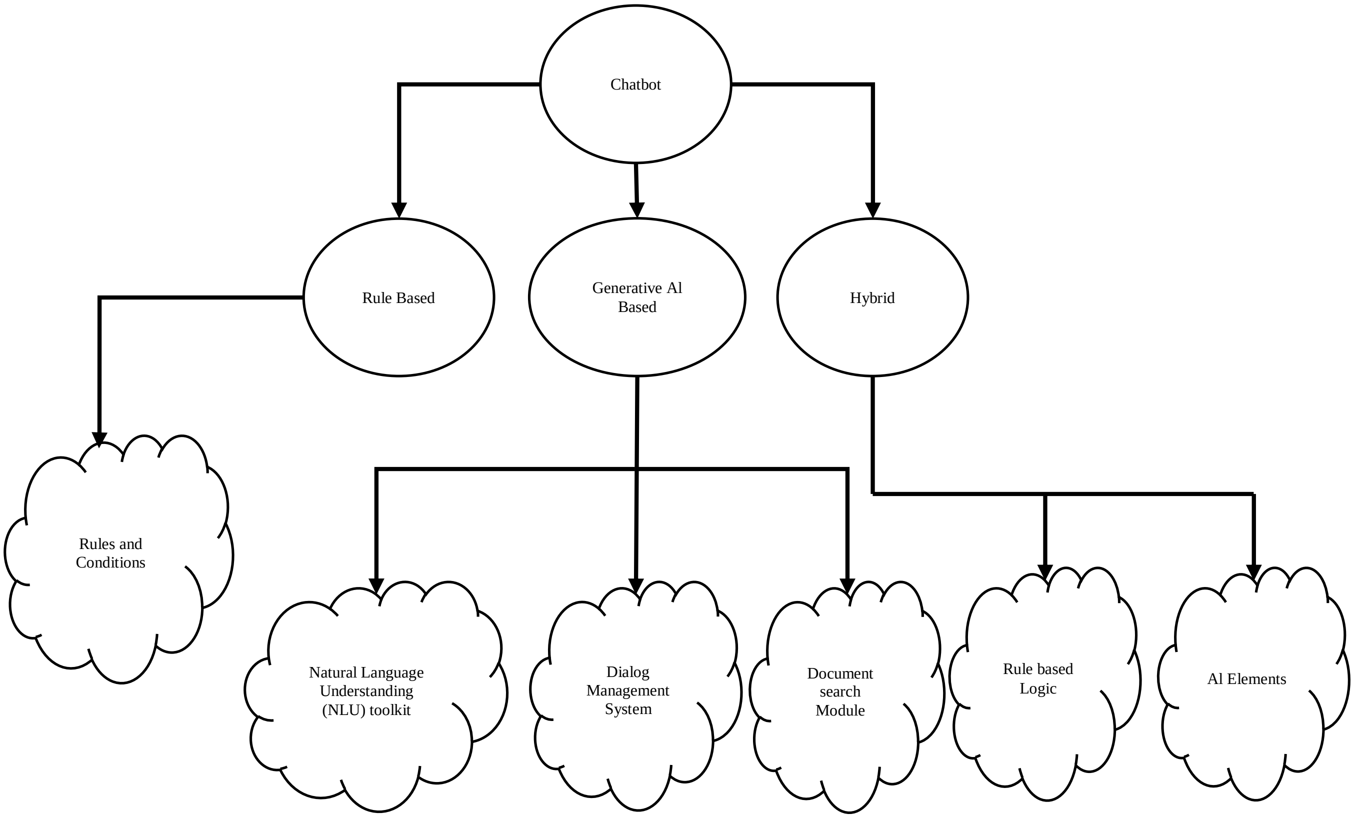

Fig. 3 General overview of chatbot

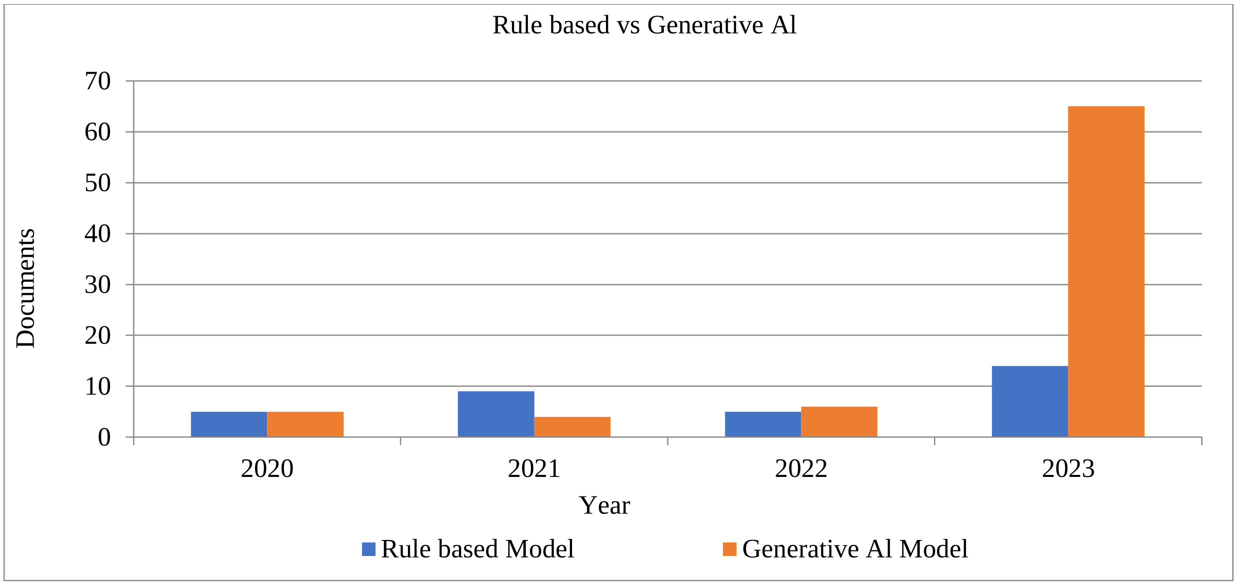

Fig. 4 Rule-based vs Generative AI model AI model research papers in Scopus with keywords "rule-based" and "generative" chatbots

Table 3. Test AI chatbots of 2024

Sl. No	Chatbot	Developed By	Purpose	Domain
1	ChatGPT	Microsoft backed start-up Open AI	Best All-Rounder	General
2	Google Gemini	Google	Best ChatGPT Alternative	General
3	Claude	Anthropic	Best for Large Inputs/Document	Specific

			Review	
4	Grok	Elon Musk’s company xAI	Best for Entertaining Conversations	Social/ Companion
5	Write sonic/ Chat sonic	Write sonic	Best for Content Creation	Specific
6	Copilot (Bing chat)	Microsoft	Best for chatbot + Web Search	General
7	Perplexity AI	Perplexity.ai	Best for Research	Specific
8	Pi	Inflection	Best Personal AI	Social/ Companion
9	Personal AI	Human Labs Inc	Best Personal Assistant	Social/ Companion
10	Poe	Quora	Best Chabot Aggregator	Specific
11	YouChat	You.com	Best Copilot Alternative	General
12	Character AI:	Character AI	Great Fun & Generates Images	Specific

However, this model type is not proofed with grammatical and typographical errors in user input. If the user types the query incorrectly, the rule identification is unsuccessful. It disregards the user's past and present interactions. Using information from both recent and past messages, the Generative Model creates fresh answers to user inquiries. It uses machine learning and deep learning algorithms to create responses, making it more akin to a human chatbot. This is a challenging model to build and train. As seen in Table 3, nearly all of the newest chatbots available today operate using generative models, and the technology used is artificial intelligence [31]. Figure 3 illustrates that in 2020, the number of research articles for both models was the same. In 2021, however, the rule-based model had more papers than the other model; in 2022, the number of papers for the generative model increased; and in 2023, the number of research papers on generative AI dramatically increased.

2.5. Report Findings

This stage involves compiling all of the research articles that have been gathered and examined, as well as the algorithms and evaluation metrics used with respect to the chatbot. Data is summarized in this phase and presented for different scenarios where most of the paper's choice of algorithm is completely independent of the context, and expert evaluation is preferred over automatic evaluation. Sameera A. et al. [8] various algorithms used in this paper include Parsing, Pattern Matching, and AIML, where the user query is split into patterns or tokens. Then, it is matched with the database if the matched response is provided to the user. Suprita Das et al. [9] state in order to extract meaning from natural language, it is important to identify the purpose of the text or sentence. In this paper, the technique proposed for this is the Dialogue Act (DA) recognition technique, which is based on the user query to conclude whether the query is a question, suggestion, command or offer. Muhammad Yusril Helmi Setyawan et al. [10] propose a classification method for intent identification. Two algorithms used by the researchers for this purpose are the Naïve bayes classification algorithm and the logistic regression algorithm, which classifies user’s messages into predefined categories or queries. Albert Verasius Dian Sano et al. [11] various techniques used involve the Hierarchical clustering method, Agglomerative Nesting (AGNES), and Euclidean distance, where similar user messages are grouped together to classify them according to their respective intents. This chatbot is specifically designed for tourists visiting Indonesia. Praveen Kumar et al. [12] use a Deep Neural Network to design a chatbot - An excessive number of layers might unnecessarily raise the complexity and lower the accuracy of a basic activity.

Dijana Peras et al. [13] mention various chatbot Evaluation Metrics using Precision, Recall, etc. Laila Hidayatin et al. [14] provide an Evaluation technique for Chatbot Applications using Term Frequency and Inverse Document Frequency (TF-IDF) and Query expansion with cosine similarity where the frequency of a word is a highlight in the user query to identify the relevance of the query. Mohit Jain et.al [15] state analysis can be either Quantitative or Qualitative Data Analysis. Joao Sedoc et al. [16] is a tool that includes created datasets and curated evaluation datasets with human-annotated and automated baselines. Algorithms used involve Seq2Seq, Neural Conversational Model (NCM), and Dialogue Breakdown Detection (DBDC). Wari Maroengsit et al. [17] involve Natural Language processing algorithms, Pattern Matching, Parsing, Intent classification, Dialogue Planning, and Long short-term memory, a technique where a chatbot model is trained with a dialogue dataset and then tested with an algorithm. Qingtang Liu et al. [18] use the K means algorithm, Deep QA and Domain-specific Knowledge Base. Contributions of this paper include Developing a domain-specific chatbot, and DOG Deep QA is assessed. Siddhi Pardeshi et al. [19] stated that the Hybrid Emotion Interference Model (HEIM) provides better results for large datasets, which means a hidden emotion in user query is identified, which in turn contributes to providing better responses. Algorithms used in this research are NLP, Pattern Matching Algorithm, Naïve Bayes Algorithm, Sequence to Sequence Model (seq2seq), Hybrid Emotion Interference Model (HEIM), and Long Short-Term Memory (LSTM). Eleni et al. [1] Chatbot classification categories rely on

Knowledge Domain, Service provided Goals, Input processing and response generation method and Permissions provided. Satyendra Praneel Reddy Karri et al. [21] here, developers do not need to write chatbot responses manually. Using a sound NLP system, the chatbot can provide smart answers. Algorithms used include Bag of Words, Seq2seq, and Beam Search Decoding. Jacky Casas et al. [22] explain Chatbot Evaluation methods based on Human Centered Computing. Evaluation methods are discussed. Nithuna S et al. [23] Algorithms used in this paper include Seq2Seq, Artificial Intelligent Algorithms, Natural Language processing algorithms, and Deep Neural Networks.

Shih-Hung Wu et al. [24] explain Automatic Evaluation by learning human evaluation with BERT. Dialogue evaluation currently relies on human judges, who are generated by a generative dialogue system, to decide on the quality of the generated text. Vijayaraghavan V et al. [25] in their paper discuss various algorithm-based inspection techniques like Naïve Bayes, Support vector machines, Deep Neural Networks, Markov chains, and Natural Language processing to assess the performance of chatbots.

They suggest cross-validation as the most suitable testing procedure for these algorithms. That is splitting the dataset into training and testing data. Caldarini G. et al. [26] explain different Machine Learning algorithms and Deep Learning Techniques. Xu Han et al. [27] state for evaluation, various models include Natural Language Generation, Large Language Models (LLM), which can understand natural languages and respond to queries just like humans, Dialogue Act (DA), and Random Forest. Ganesh Reddy Gunnam et al. [28] assess the performance of Cloud-Based Heterogeneous Chatbot Systems using Natural Language Processing and Multimodal interactions. Daniel Escobar et al. [29] use Parallel Convolutional Networks (PCNN), Word Vec, BERT and BETO; all these techniques do the job of classifier on input to categorize intents. Automatic evaluation, combined with user satisfaction and performance metrics, could deliver more trustworthy results compared to human evaluation, which otherwise could be biased.

2.5.1. Datasets Used

The most popular datasets for employing deep learning techniques to train chatbots are covered in this section. Closed-domain datasets and open-domain datasets are the two basic types of datasets that are employed. A closed-domain dataset has been specifically created for a given domain and scope; for example, a Twitter dataset with all its questions is not publicly accessible and is not frequently utilized in various research publications. The Open Domain Datasets are the most widely utilized datasets by researchers that are freely available to anyone. For example, the WikiQA corpus is a publicly available question-and-answer pair. Additional datasets fall under the categories of assessment and training datasets. Where the training dataset is used to train the model to make itself aware of the relationship between the inputs and their outputs, and the evaluation dataset is used to evaluate the model's performance for which it is trained.

While the dialogue breakdown detection dataset is utilized for evaluation, the Cornell and question-answering datasets are used for training. Cornell corpus consists of rich collections of fictional conversations from raw movie scripts, and its metadata comprises genres, IMDB rating, release year, etc. The metadata of the question-answering dataset presented by AliMe systems comprises a Question log, highly frequent entities, highly frequent questions, etc. Dialogue Breakdown Detection Challenge (DBDC) corpus is to detect a situation where users cannot proceed further with computer conversation. These include three tasks: Dialogue Breakdown Detection: Breakdown is detected in this phase, Error category Classification: Error categories to describe the causes of breakdown, Recovery response generation: System must be able to provide a new response by recovering or correcting the reason for breakdown.

Table 4. Summary of commonly use dataset in chatbots

Dataset	Type of Dataset	Content	Phrases	Source
Cornell	Training	Scripts of raw movie	304713	[21]
Question Answer	Training	AliMe System	9,164,834	[30]
DBDC	Evaluation	Dialogue breakdown	NA	[16]

2.5.2. Evaluation

Researchers focused on the design and efficient operation of chatbots in the early stages of their development. The examination of chatbots was overlooked. The primary cause was the lack of organization or structure in the parameters used for evaluating chatbots. However, it has been noted that studies about chatbots and their evaluation have been more prevalent in recent years. According to this study, the evaluation of chatbots gained attention starting in 2017, and the number of papers pertaining to this topic has been steadily rising ever since as shown in Figure 5.

2.5.3. Performance Assessment of a Chatbot

Chatbots must undergo extensive testing, validation, and verification to prevent them from failing throughout the procedure. Chatbots need to be able to manage and handle situations, even in the event of failures. Testing algorithms is a potential remedy for these kinds of issues. Chatbot testing can be carried out in two ways: either by observing the chatbot's output and assessing its performance or by comprehending and looking into the inner workings of the chatbot and the different algorithms under consideration [29]. The chatbot's performance can be evaluated using quantitative or qualitative measures.

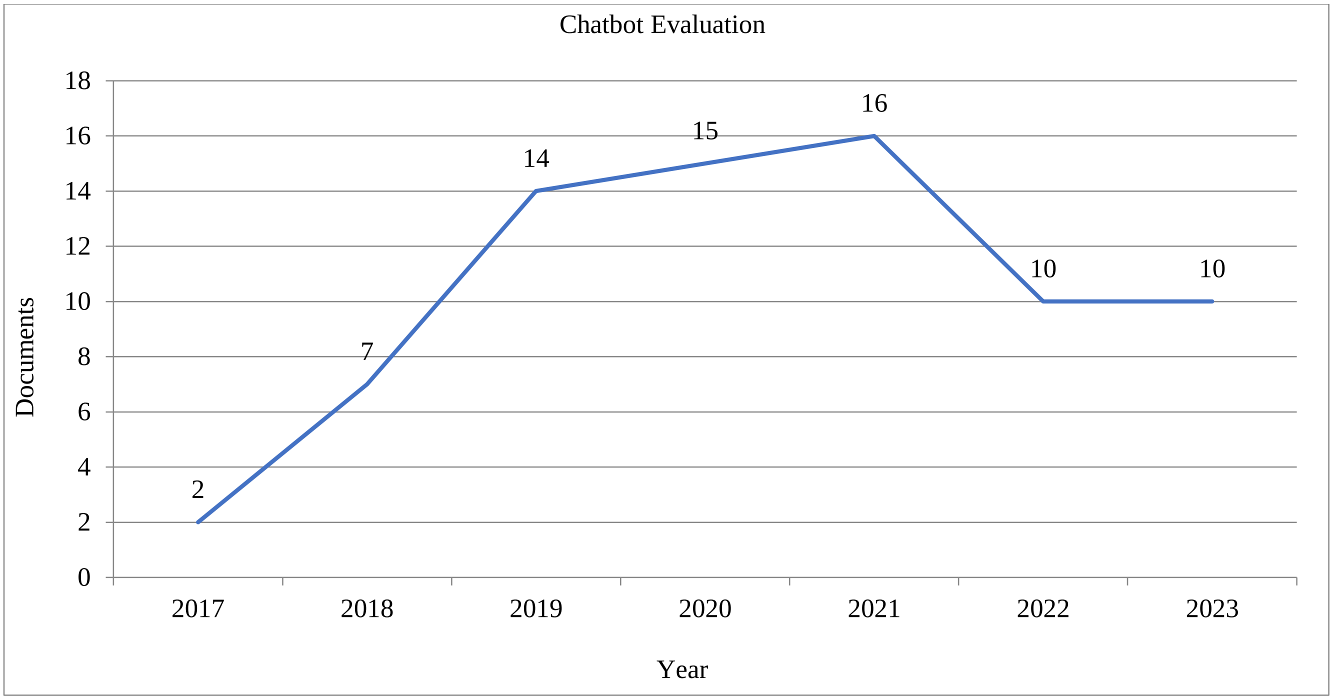

Fig. 5 Results of search, from 2017 to 2023 in Scopus for the keywords "chatbot evaluation" or "chatbot evaluation"

Table 5. Defined research question for chatbot performance evaluation

Question No	Question	Criteria of Evaluation
Q5	What are the current testing approaches followed by monitoring the output of the chatbot?	Quality of the response, relevance, completeness, accuracy and context. Feedback from the user is the primary concern Table 8.
Q6	What are the testing approaches for studying the inner functionality of the chatbot?	Techniques of algorithm inspection Table 6.

In general, evaluations fall into the following categories:

- First Meta-Evaluation: Utilizes measurements for effectiveness, satisfaction, and efficiency.
- Second Meta-Evaluation: Methods of Assessment
  a) Content Assessment
  b) User Contentment
  c) Assessment of Function
- Conversational AI prospects
  a) The viewpoint of the user
  b) From the standpoint of information retrieval: How quickly and accurately does a chatbot answer a user's question.
  c) From a linguistic perspective: Points for the quantity, quality, relationships, and mannerisms of the discussion.
  d) From an AI standpoint: Can the chatbot pass the Turing test?
- HCI perspectives: In terms of human assessment, these include the chatbot's functionality, intellect, personality, and interface.

Table 6. Defined research question in analysis of evaluation in chatbots

Question No	Question	Criteria of Evaluation
Q7	What are the different methods used to evaluate a chatbot?	Evaluation can be conducted based on the content, task and user satisfaction. Table 7
Q8	What are the techniques used to evaluate a chatbot based on data?	Data can be categorised into quantitative data and qualitative data. Table 8

Table 7. Chatbot evaluation methods

Sl. No.	Evaluation Method	Types	Merits	Demerits	Source
1	Content Evaluation: Collection of methods to evaluate chatbots	1. Automatic Evaluation	• Quick evaluation • Cost friendly	• Not accurate in every field • Not trustworthy	[16, 17, 26, 22]
		2. Expert Evaluation	• Trustworthy • Accurate in every field	• Expensive • Human evaluation can be bias	[16, 17, 22, 26]
2	User Satisfaction: Users' interaction with the chatbot and their satisfaction rate.	1. Session Level	• Rate the entire session	• Entire session chatbot may not be accurate during user interaction.	[17, 22]
		2. Turn Level	• Evaluate each response from the chatbot.	• Every answer of the chatbot to users' queries may not be right.	[17, 22]
3	Functional evaluation: Evaluation based on goal/task	1. Usage statistics	• Usually, it indicates better performance.	• It may not always be true.	[17, 22]
		2. Building blocks for chatbots	• Evaluation can be conducted both intermediate or final product evaluation	• Multiple times, evaluating the product may be costly.	[17, 22]

Data can also be used to categorize evaluations.

Table 8. Chatbot evaluation based on analysis of data

Sl. No	Data Analysis	Measures	Sources
1	Quantitative Data analysis	• Task Completion Rate (TCR) • Number of turns • Total Time	[13, 15]
2	Qualitative Data Analysis	• Functionality • Conversational Intelligence • Personality and Interface	[13, 15]

3. Trust in Chatbots

Humans place more trust in computers than humans because we as humans believe we are prone to make mistakes, unlike machines Figure 6. But on the same ground, chatbots are expected to behave more like humans in terms of features, behavior, relation, manners etc. Today, in almost all fields' people trust chatbots may, maybe healthcare, the banking sector, etc. Chatbot performance can also be evaluated using the following quantitative metrics, which deal with numerical value or statistical data:

- Mean Squared Error (MSE): The smaller the error, the better the chatbot.
- Area Under the Receiver Operating Characteristic
  - ➢ Curve (AUC-ROC): Visually represents tradeoffs between True positive rate and False positive rate at varying thresholds. AUC value ranges between 0 to 1.
- Confusion Matrix: Shows how well the model performs by predicting correct and incorrect predictions. Metrics under it comprise of:
  - ➢ Accuracy: Overall correct predictions of the model.
  - ➢ Recall: Actual Positive classes identified by the model.
  - ➢ Precision: Models positive predictions.
  - ➢ F1 score: Combines results of Precision and Recall.

- The efficiency of a chatbot also relies on the choice of algorithms in combination with optimizers. The results obtained from these metrics could be a deciding factor in measuring the trustworthiness of chatbots.

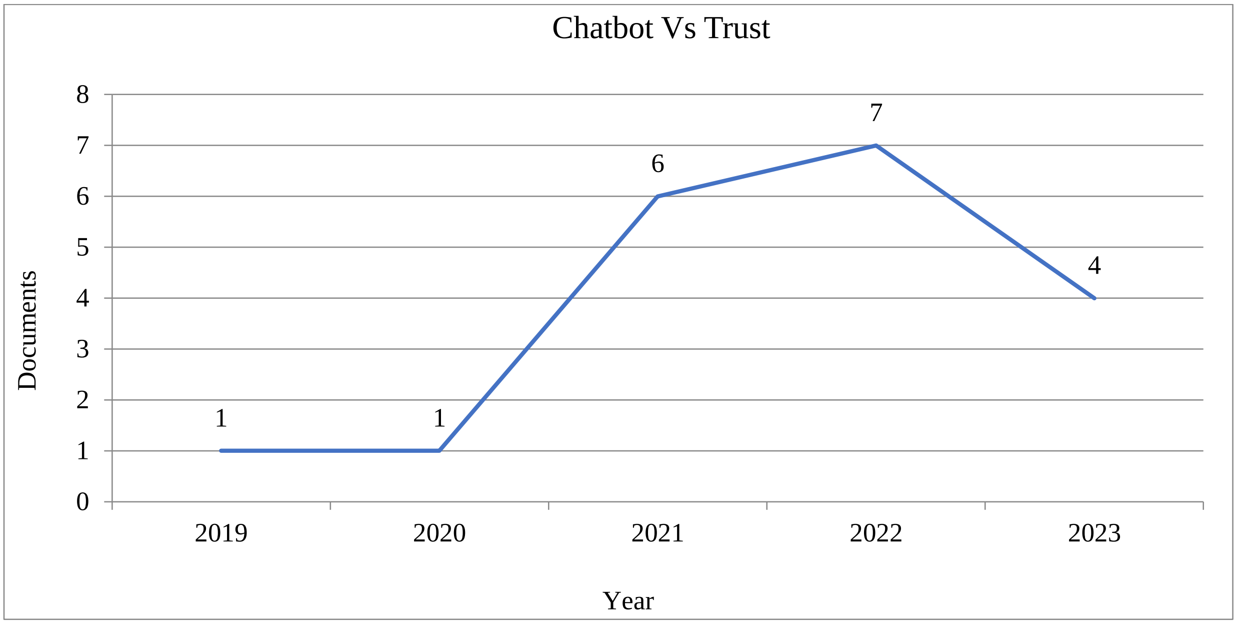

Fig. 6 Results of search, from 2019 to 2023 in Scopus for the keywords “chatbot trust” or “chatbots trust”

4. Discussion

As per the study, it can be concluded that the choice of trustworthy metrics is independent, irrespective of the context in which the chatbot is used. Quantitative measure accuracy cannot be considered the best metric to measure trust for chatbots used in messaging apps. However, user satisfaction can be considered the best choice to measure trust in messaging apps, and vice versa; banking apps may consider the precision of responses as the best metric to measure the efficiency of a chatbot.

The choice of algorithm in chatbot inspection depends on the dataset used, and different algorithms hold good for different kinds of data. Performance measurement of chatbots through human evaluation relies on qualitative measures like user satisfaction usage statistics, which may be biased and cannot be considered a trustworthy metric, whereas automatic evaluation could give better results based on quantitative measures like accuracy, precision, etc. Chatbot efficiency could be proved by using figures and statistics and automatic evaluation.

5. Conclusion

The paper presented identifies a study of different states of methods used in chatbots, variants on chatbots, and applications of chatbots. This research study provides a detailed overview of the work published on chatbots, ranging from 2015 to 2024. Throughout the paper, it has been tried to answer all possible questions concerning chatbots by discussing starting with its history, application, possible variants, existing limitations, and various chatbot technologies and methods and executing a comparative, summarized study of various works based on their existing works. Key findings and observations derived from the study are that the chatbot has no universal evaluation approach due to time and financial constraints. The use of a limited amount of training data leads to accuracy, which may not be accurate. The chatbot is tested specifically, but the results remain uncertain in other domains. The lack of quantitative metrics is a major drawback in evaluation metrics.

Ambiguous statements have posed problems when given as input to chatbots, which may fail to respond meaningfully. The user makes use of a lot of abbreviations and slang words and communicates differently at different times. It is difficult to compare two sentences only with their content; the same content could have two different meanings. Still, there is dependence on human evaluation, which requires replacing it with automatic evaluation. The user satisfaction rate is considered one of the evaluation metrics of a chatbot, but this metric is not the correct and accurate method for measuring the efficiency of a chatbot because every answer of a chatbot to a user’s query may not be right. Even though chatbot is considered more trustworthy than humans, when it comes to evaluation, it is considered that automatic evaluation is not trustworthy. Still, the trust relies on expert evaluation, even though humans could be biased in their decisions. Usage statistics are considered one of the factors for evaluation, but they cannot be considered as a deciding metric in the evaluation of chatbots. Every user is different from others, and the time taken by every user varies indifferently.

References

[1] Eleni Adamopoulou , and Lefteris Moussiades “An Overview of Chatbot Technology”, *Artificial Intelligence Applications and Innovations 16th IFIP WG 12.5 International Conference*, Neos Marmaras, Greece, pp. 373-383, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Rohit Tamrakar, and Niraj Wani, “Design and Development of CHATBOT: A Review,” *International Conference On “Latest Trends in Civil, Mechanical and Electrical Engineering”,* Bhopal, India, pp. 1-15, 2021. [Google Scholar]

[3] A.M. Turing, “I- Computing Machinery and Intelligence,” *Mind*, vol. LIX, no. 236, pp. 433-460, 1950. [CrossRef] [Google Scholar] [Publisher Link]

[4] Joseph Weizenbaum, “ELIZA- A Computer Program for the Study of Natural Language Communication between Man and Machine,” *Comnutnieations of the ACM*, vol. 9, no. 1, pp. 36-45, 1966. [Google Scholar] [Publisher Link]

[5] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf, “Artificial Paranoia,” *Artificial Intelligence*, vol. 2, no. 1, pp. 1-25, 1971. [CrossRef] [Google Scholar] [Publisher Link]

[6] Richard S. Wallace, *The Anatomy of A.L.I.C.E*, Parsing the Turing Test Philosophical and Methodological Issues in the Quest for the Thinking Computer, Springer, Dordrecht, pp 181-210, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[7] P. Costa, “Conversing with Personal Digital Assistants: On Gender and Artificial Intelligence,” *Journal of Science and Technology of the Arts*, vol. 10, no. 3, pp. 59-72, 2018. [Google Scholar] [Publisher Link]

[8] Sameera A. Abdul-Kader, and John Woods “Survey on Chatbot Design Techniques in Speech Conversation Systems” *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, pp. 72-80, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[9] Suprita Das, and Ela Kumar, “Determining Accuracy of Chatbot by Applying Algorithm Design and Defined process,” *4th International Conference on Computing Communication and Automation*, Greater Noida, India, pp. 1-6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Muhammad Yusril Helmi Setyawan, Rolly Maulana Awangga, and Safif Rafi Efendi, “Comparison of Multinomial Naive Bayes Algorithm and Logistic Regression for Intent Classification in Chatbot,” *IEEE International Conference on Applied Engineering*, Batam, Indonesia, pp. 1-5, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Albert Verasius Dian Sano et al., “The Application of AGNES Algorithm to Optimize Knowledge Base for Tourism Chatbot,” *IEEE International Conference on Information Management and Technology*, Jakarta, Indonesia, pp. 65-68, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Praveen Kumar et al., “Designing and Developing a Chatbot Using Machine Learning,” *IEEE International Conference on System Modeling and Advancement in Research Trends*, Moradabad, India, pp. 87-91, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Dijana Peras, “Chatbot Evaluation Metrics-Review Paper,” *36th International Scientific Conference on Economic and Social Development - Building Resilient Society*, Zagreb, pp. 89-97, 2018. [Google Scholar] [Publisher Link]

[14] Laila Hidayatin, and Faisal Rahutomo, “Query Expansion Evaluation for Chatbot Application,” *IEEE International Conference on Applied Information Technology and Innovation*, Padang, Indonesia, pp. 92-95, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohit Jain et al., “Evaluating and Informing the Design of Chatbots,” *Proceedings of the Designing Interactive Systems Conference*, Hong Kong, China, pp. 895 - 906, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] João Sedoc et al., “ChatEval: A Tool for Chatbot Evaluation,” *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, pp. 60-65, 2019. [Google Scholar] [Publisher Link]

[17] Wari Maroengsit et al., “A Survey on Evaluation Methods for Chatbots,” *Proceedings of the 7th International Conference on Information and Education Technology*, Aizu-Wakamatsu, Japan, pp. 111-119, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] Qingtang Liu et al., “CBET: Design and Evaluation of a Domain-Specifc Chatbot for Mobile Learning,” *Universal Access in the Information Society International Journal*, vol. 19, pp. 655-673, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Siddhi Pardeshi et al., “A Survey on Different Algorithms used in Chatbot” *International Research Journal of Engineering and Technology*, vol. 7, no. 5, pp. 6092- 6098, 2020. [Google Scholar] [Publisher Link]

[20] Anirudh Khanna et al., “A Study of Today’s A.I. through Chatbots and Rediscovery of Machine Intelligence,” *International Journal of u- and e- Service, Science and Technology*, vol. 8, no. 7, pp. 277-284, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[21] Satyendra Praneel Reddy Karri, and B. Santhosh Kumar, “Deep Learning Techniques for Implementation of Chatbots,” *IEEE International Conference on Computer Communication and Informatics*, Coimbatore, India, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] Jacky Casas et al., “Trends & Methods in Chatbot Evaluation,” *Companion Publication of the International Conference on Multimodal Interaction*, Virtual event, Netherlands, pp. 280-286, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] S. Nithuna, and C.A. Laseena, “Review on Implementation Techniques of Chatbot,” *IEEE International Conference on Communication and Signal Processing*, Chennai, India, pp. 0157-0161, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Shih-Hung Wu, and Sheng-Lun Chien, “Learning the Human Judgment for the Automatic Evaluation of Chatbot,” *Proceedings of the 12th Conference on Language Resources and Evaluation*, Marseille, France, pp. 1598-1602, 2020. [Google Scholar] [Publisher Link]

[25] V. Vijayaraghavan, Jack Brian Cooper, and J. Rian Leevinson, "Algorithm Chatbot Inspection for Chatbot Performance Evaluation," *Third International Conference on Computing and Network Communications*, *Procedia Computer Science*, vol. 171, pp. 2267-2274, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry, "A Literature Survey of Recent Advances in Chatbots," *Information*, vol. 13, no. 1, pp. 1-22, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Xu Han et al., "Democratizing Chatbot Debugging: A Computational Framework for Evaluating and Explaining Inappropriate Chatbot Responses," *Proceedings of the 5th International Conference on Conversational User Interfaces*, Eindhoven, Netherlands, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Ganesh Reddy Gunnam et al., "Assessing Performance of Cloud-Based Heterogeneous Chatbot Systems and A Case Study", *IEEE Access*, vol. 12, pp. 81631-81645, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Daniel Escobar-Grisales, Juan Camilo Vasquez-Correa, and Juan Rafael Orozco-Arroyave, "Evaluation of Effectiveness in Conversations Between Humans and Chatbots Using Parallel Convolutional Neural Networks with Multiple Temporal Resolutions Multimedia Tools and Applications," *Multimedia Tools and Applications*, vol. 83, pp. 5473-5492, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Abbas Saliimi Lokman, and Mohamed Ariff Ameedeen, "Modern Chatbot Systems: A Technical Review," *Proceedings of the Future Technologies Conference*, vol. 2. pp. 1012-1023, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[31] Aaron Drapkin, 13 Best Free and Paid AI Chatbots in 2024: ChatGPT, Gemini & More, 2024. [Online] Available: https://tech.co/news/best-ai-chatbots

[32] Juliette Mattioli et al., "An Overview of Key Trustworthiness Attributes and Kpis for Trusted Ml-Based Systems Engineering," *AI and Ethics*, vol. 4, no. 1, pp. 15-25, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Nitin Rane, Saurabh Choudhary, and Jayesh Rane. "Artificial Intelligence (AI), Internet of Things (IoT), and Blockchain-Powered Chatbots for Improved Customer Satisfaction, Experience, and Loyalty," *SSRN Electronics Journal*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Zuhal 'Alimul Hadi et al., "The Influence of Transparency, Anthropomorphism, and Positive Politeness on Chatbots for Service Recovery in E-Health Applications," *Cogent Social Sciences*, vol. 10, no. 1, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Jijie Zhou, and Yuhan Hu, "Beyond Words: Infusing Conversational Agents with Human-like Typing Behaviors," *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, Luxembourg, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Cecylia Borek, "*Comparative Evaluation of LLM-Based Approaches to Chatbot Creation*," Master's Thesis, Tampere University, pp. 1-64, 2024. [Google Scholar] [Publisher Link]

[37] Asad Ali, "Assessing AI Chatbots through Meta-Analysis of Deep Learning Models," *EasyChair Preprint*, pp. 1-10, 2024. [Google Scholar] [Publisher Link]

[38] Xiaojie Wang et al., "A Survey on Trustworthy Edge Intelligence: from Security and Reliability to Transparency and Sustainability," *IEEE Communications Surveys and Tutorials* (Early Access), pp. 1-1, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Muhammad Farrukh Shahzad et al.,"Assessing the Impact of AI-Chatbot Service Quality on User E-Brand Loyalty through Chatbot User Trust, Experience and Electronic Word of Mouth," *Journal of Retailing and Consumer Services*, vol. 79, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[40] Jesús Sánchez Cuadrado et al., "Automating the Development of Task-Oriented LLM-Based Chatbots," *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, Luxembourg, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]