Evaluation Metrics and Ethical Concerns in the Design of Adaptive AI Chatbots

Akrivi Krouska
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, Greece
akrouska@uniwa.gr

Christos Troussas
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, Greece
ctrouss@uniwa.gr

Phivos Mylonas
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, Greece
mylonasf@uniwa.gr

Ioannis Voyiatzis
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, Greece
voyageri@uniwa.gr

Cleo Sgouropoulou
Department of Informatics and Computer Engineering
University of West Attica
Egaleo, Greece
csgouro@uniwa.gr

***Abstract*— The proliferation of large language models in recent years has led to the widespread development of AI chatbots. AI chatbots are smart systems that enable conversations with users using natural language. Chatbots are used to automate tasks, offer real-time assistance, and enhance user experiences. Their increasing use in many domains, such as customer service, education and healthcare support, makes their reliable evaluation essential. To this direction, this paper presents a literature review on evaluation frameworks for adaptive AI chatbots, highlighting the trends, gaps, and future research directions. It focuses on four critical dimensions of chatbot assessment: personalization and adaptability, information accuracy, response adequacy and speed, as well as bias and ethical concerns. The comparative analysis of related studies shows the importance of applying multidimensional evaluation approaches that integrate both technical features and user-centered metrics. Furthermore, it identifies major challenges, such as the lack of standardized evaluation criteria and the need for transparency and ethical accountability in chatbot deployment.**

Keywords*— *Conversational Systems, Ethical AI, Evaluation Frameworks, Large Language Models, Personalization, Usability Metrics

I. INTRODUCTION

The rapid growth of digital technologies, and especially of artificial intelligence (AI), has given rise to AI chatbots [1]. These are conversational systems that use intelligent techniques to mimic human dialogue, enabling users to interact with computers through natural language [2]. Today, chatbots are widely used across various sectors, such as customer service, healthcare, education, and e-commerce, where they automate tasks, provide instant responses, retrieve useful information, and personalize interactions based on user behavior [3-6]. Due to their ability to simulate human-like and real-time conversations, chatbots have become a valuable tool in many aspects of daily life.

As chatbots' use become more widespread, the need for systematic evaluation is more essential. In particular, it is important to ensure that these systems function accurately, efficiently, and ethically in order users to trust them [7-9]. Evaluation not only identifies weaknesses but also guides improvements and ensures that chatbot behavior aligns with user expectations and societal norms [9-10]. However, the dynamic and human-centric nature of chatbot interactions often challenges traditional evaluation methods. This has led to the development of specialized frameworks that assess multiple dimensions of chatbot performance, including personalization, information accuracy, responsiveness, and ethical behavior [11]. These frameworks combine technical metrics with user-centered approaches to provide a more holistic understanding of how chatbots perform in real-world settings [10].

Some studies in this field focus on personalization and user adaptation as core evaluation dimensions, highlighting the importance of adjusting chatbot behavior to individual preferences and personality traits [12]. Others propose quantitative performance metrics, such as answerability, task completion, or error correction, as tools for measuring chatbot effectiveness in real time [7, 13]. Many studies also explores methodological frameworks and taxonomies for evaluating conversational agents across various contexts, emphasizing on usability, responsiveness, and content quality [1, 9-11]. Finally, several researches investigate the ethical and social implications of chatbot use, introducing frameworks that assess transparency, fairness, data privacy, and accountability [8, 14].

In view of the above, this paper presents a comparative analysis of established evaluation frameworks of intelligent chatbots, aiming to highlight the most relevant concepts, methods, and emerging challenges. As such, it provides an overview of the key evaluation dimensions, namely personalization, adaptability, response accuracy, speed, bias, and ethical concerns. The scope of this comparative analysis is to explore how chatbot performance is currently assessed, what gaps remain, and which areas demand further exploration. Moreover, this review aims to identify the most commonly used evaluation frameworks, assess their effectiveness, and investigate what makes each method valuable and reliable. Based on the findings, this study proposes directions for an integrated AI chatbot evaluation framework that could support the development of more robust, fair, and human-centered conversational AI systems.

II. RESEARCH METHODOLOGY

This study follows the five-step framework for scoping reviews proposed by Arksey and O'Malley (2005) [15], which provides a systematic approach to organizing and synthesizing the existing literature. The five stages include:

1) identifying the research questions,
2) identifying relevant studies,
3) study selection,
4) charting the data, and
5) collating, summarizing, and reporting the results.

Regarding the research questions of this study, they were designed to explore different aspects of chatbot evaluation, from personalization and performance to ethical considerations. Thus, the research questions defined are the following:

- RQ1: How is a chatbot's personalization and adaptability to individual user preferences and needs evaluated?
- RQ2: What methods are used to assess the accuracy of the information provided by AI chatbots?
- RQ3: How is the adequacy and response speed of AI chatbot answers evaluated in current research?
- RQ4: How are ethical considerations addressed in chatbot evaluation models?

Afterwards, the search criterion were defined in order to find the most relevant studies. As such, keywords and phrases that include both broad and specific terms derived from each research question, were used, like "AI chatbot evaluation", "chatbot evaluation framework", "AI chatbot personalization assessment", "AI chatbot performance evaluation", "evaluation metrics for chatbots", "ethical considerations in conversational AI" etc. This strategy helped find a diverse range of studies across technical, methodological, and ethical dimensions of chatbot evaluation. The search engines used in this study were Scopus, Google Scholar, and the IEEE Xplore digital library, since they provide access to a wide range of peer-reviewed publications.

After the initial identification of potentially relevant studies, a systematic screening process was applied to determine which publications would be included in the final review. The selection was based on a set of predefined inclusion and exclusion criteria designed to ensure both the quality and relevance of the chosen literature. Thus, the studies had to be written in English, published between 2018 and 2024, and provided clear methodological details or theoretical contributions relevant to the research questions. From this step, around 35 studies was selected for review in this work.

The selected studies were systematically organized and classified according to the research questions. Each paper was reviewed to extract relevant information regarding the evaluation approach they used.

Finally, the extracted data were synthesized to highlight common practices, identify gaps in the literature, and suggest directions for future research. The findings were interpreted in relation to the research questions.

III. Comparative Analysis

A. Personalization and adaptability

Evaluating how well chatbots personalize content and adapt to individual users is essential for their acceptance and adoption in everyday life. To this direction, it is needed to clarify what is being adapted, namely tone, content, recommendations, dialog strategy, and for whom, like user traits, preferences, current goal and context. Thus, there are studies that combine technical measures with human-experience outcomes [3, 12, 16-17], while others that use frameworks aligned with ISO usability dimensions, such as effectiveness, efficiency, and satisfaction, augmented with constructs such as engagement, empathy, and task success for different user segments [9, 18-19].

Collaborative filtering approaches have been employed to tailor recommendations, accelerating information retrieval and reservation tasks while improving user satisfaction, but they remain dependent on large-scale historical datasets and raise privacy concerns [3, 20-21]. Comparative studies between chatbot-based and menu-based interfaces reveal that adaptability to user autonomy and cognitive load plays a critical role in perceived usability, with chatbots performing better in exploratory tasks but sometimes increasing mental effort for goal-oriented users [22].

In applied service domains, natural language processing combined with neural network architectures has enabled real-time personalization and rapid response generation, reducing operational costs while improving task success rates [23]. Moreover, advanced machine learning models perform well in understanding user requests and adjusting their responses properly, and thus making interactions over time more personalized [7, 24].

Overall, the literature highlights that effective evaluation requires a multi-method approach that combines both objective task performance and subjective user experience, in order to ensure that personalization mechanisms align with user preferences and needs.

B. Information accuracy

It is crucial to ensure that AI chatbots provide accurate information in order to trust and use them effectively, especially in sensitive domains like healthcare, education, and financial services. Multiple studies have shown that chatbots that incorporate Large Language Models (LLM) exhibit variability depending on the topic domain, language, and source of information. For instance, in [5], the authors conducted a comparative analysis between chatbot-generated answers and responses from human medical experts, revealing that while LLMs demonstrated high accuracy rates in standardized, fact-based questions, they showed reduced consistency in open-ended queries, with statistically significant deviations exceeding 15%. Similarly, in [4], the authors investigated the ability of chatbots to verify political information and found accuracy rates of 72% for ChatGPT and 67% for Bing Chat, highlighting the difference in performance across platforms.

Researchers have explored many ways to detect and fix errors in order to boost chatbot accuracy. In [13], the authors demonstrated that the integration of reinforcement learning and user feedback loops reduced error rates by 28% and increased overall chatbot accuracy by up to 20%. This confirms that continuous model refinement can significantly

enhance output quality. In a related study [25], the authors applied cosine similarity to measure semantic closeness between chatbot-generated and reference answers, providing a quantitative approach to assessing retrieval accuracy.

Another major challenge is the management of misinformation and AI hallucinations, where LLMs generate fabricated or misleading content [26-28]. In [26], the authors introduced a multi-model consensus strategy that compared outputs from different chatbots, reducing hallucination rates by 23% and improving answer consistency by 17%. This approach indicates that cross-model verification can be a powerful safeguard against inaccurate outputs.

C. Adequacy and response speed of AI chatbot answers

The literature review showed that researchers use a wide range of criteria and methodologies for evaluating the adequacy and response speed of intelligent chatbots, including both technical performance and the overall user experience [29-30].

Firstly, it is examined the ability of chatbots to deliver correct and contextually appropriate answers to user queries. In [7], the authors introduced the concept of "answerability" as a quantitative measure of response quality, assigning scores from −1 to +1 based on correctness and completeness. Similarly, in [31], the authors developed the *Chattest* framework, which uses a fixed set of test questions to check the precision and relevance of chatbot answers in a systematic way.

Regarding usability and user satisfaction, the authors of [1] emphasized that chatbot usability involves effectiveness, efficiency, and satisfaction in achieving user goals, while in [38], the authors highlighted trust, perceived usefulness, and emotional engagement as critical determinants of a successful interaction.

Linguistic quality and anthropomorphism also emerged as key evaluation factors. Studies have highlighted the need for chatbots to be linguistically accurate, coherent, and natural in their responses [10-11]. To measure this, researchers have used a mix of quantitative tools like BLEU scores and lexical diversity metrics, as well as qualitative approaches such as user surveys, to evaluate how fluent and human-like the conversation is.

In addition, functional performance and efficiency have been assessed using operational metrics such as task completion rates, conversation duration, and frequency of required human intervention [9, 32]. From a business perspective, operational efficiency is closely tied to cost savings and the degree of automation in customer interaction workflows [9].

Finally, technological performance and system reliability have been investigated not only in terms of the chatbot's outputs but also with respect to the underlying algorithms. In [30], the authors demonstrated an open-source system enabling comparative evaluations of different chatbots through standardized technical metrics, allowing researchers and practitioners to benchmark systems in a reproducible manner.

D. Ethical considerations

The evaluation of biases and ethical considerations in AI chatbots encompasses two distinct but interrelated domains [33-35]. The first concerns biases, which are either those learned and perpetuated by chatbots or those that emerge in users as a result of interacting with them. The second addresses broader ethical issues, which can span multiple dimensions such as fairness, accountability, privacy, and transparency. In the context of bias detection, evaluation frameworks are often assessed based on the breadth of biases they can identify and their ability to determine, or at least provide evidence toward determining, the underlying causes. For ethics, the relevant evaluation criteria relate to the range of ethical concerns addressed by the framework and the extent to which causal factors are explored.

Empirical studies have examined the detection of demographic and cognitive biases in chatbot outputs using both human-centered experiments and automated interrogation of chatbot systems. In [36], the authors investigated the degree to which large language models influence human evaluators, finding that while such biases can be identified relatively easily, pinpointing the precise source within the model architecture or training data remains challenging. In [37], the authors applied structured prompts to multiple AI chatbots and categorized responses according to bias scales, highlighting the difficulty of achieving a universally accepted operational definition of bias in conversational AI. In [38], the authors advanced this line of inquiry by employing association tests to measure social bias in GPT-based models, demonstrating that a generalized evaluation framework can be adapted to detect a wide range of prejudices, though subjectivity in bias definitions limits interpretability.

On the ethical dimension, in [14], the authors applied a multi-dimensional moral framework to chatbot outputs, enabling the identification of issues across diverse ethical categories, from data privacy to harmful content generation. However, the authors noted a lack of domain-specific granularity when applying such frameworks exclusively to chatbots.

In [8], the authors conducted a comparative study of existing ethical AI frameworks in practical chatbot deployments, identifying missing elements in current approaches, particularly with respect to transparency and accountability in automated decision-making. Collectively, these studies underscore that while current evaluation methods can surface both bias-related and ethical shortcomings, limitations persist due to subjective definitions, incomplete causal mapping, and the absence of chatbot-specific ethical evaluation standards. Addressing these challenges requires not only methodological refinements but also the development of domain-specific tools tailored to conversational AI systems.

IV. Discussion

A. Trends in AI Chatbot Evaluation

The analysis of recent studies on AI chatbot evaluation reveals several important trends that define the current state of the field. In the area of personalization, there is a clear move away from traditional evaluation methods, such as the Turing Test, toward more multidimensional evaluation frameworks. These newer approaches combine both numerical and qualitative measures, including user engagement and empathy. Frameworks like SPACs and OPACs support the adaptation of chatbot responses to individual user needs, improving the quality of interaction [1, 12].

In the field of ethics, although general guidelines for AI exist [8, 14], many of them are not specifically designed for chatbots. Recent studies, however, are beginning to address this gap by focusing on the detection and reduction of bias, using clearer definitions and more objective methods [37-38].

Regarding information accuracy, modern approaches include the use of AI auditing, natural language processing techniques for semantic similarity, and real-time fact-checking tools. These help improve the consistency and trustworthiness of chatbot responses, especially in areas like healthcare where accuracy is critical [4-5, 22, 25].

Moreover, there is increasing use of combined evaluation methods that take into account both technical performance and the user experience. These include automatic measurements along with human feedback related to usability and satisfaction. Such combined methods help provide a more complete understanding of how well chatbots perform in real conditions [2, 7, 11].

B. Gaps in AI Chatbot Evaluation

In the area of personalization, many studies do not offer clear or standardized frameworks. Real-time adaptation and personality recognition remain difficult challenges, and many models do not fully consider cultural and emotional differences, which limits their ability to meet the needs of diverse users [3, 12].

In terms of ethical evaluation, there is still a lack of practical tools for detecting bias in newer chatbot models. Often, the source of bias is not clearly identified, and the definitions used can be too subjective, making evaluation inconsistent [37-38]. In addition, privacy protection is not always addressed in depth, and more work is needed to develop effective and transparent evaluation methods [14].

As for information accuracy, current evaluation methods mostly rely on structured prompts, which do not reflect how users interact with chatbots in real life. Furthermore, many systems perform poorly in languages other than English due to limited multilingual support [13, 36].

Finally, regarding efficiency, the lack of a common evaluation standard makes it difficult to compare different chatbot systems. Also, the use of subjective measures, such as user satisfaction, can lead to results that are not consistent or easy to interpret [10, 22, 31].

C. Proposed Integrated Evaluation Framework for AI Chatbots

Based on the comparative analysis presented in this study, an integrated evaluation framework designed to holistically assess AI chatbot performance is proposed. This framework adopts a multidimensional approach, which combines quantitative with qualitative metrics to provide an holistic analysis of AI chatbot performance. The proposed framework is structured based on four key dimensions: i. personalization and adaptability, ii. information accuracy, iii. response adequacy and speed, and iv. bias and ethical considerations. For each dimension, specific metrics and methods are defined, which range from semantic similarity scores and fact verification rates to bias detection measures and user satisfaction indicators. As such, the framework integrates technical capabilities along with user experience and ethical compliance. Table 1 illustrates the four dimensions of the framework.

Table 1. The Proposed Integrated AI Chatbot Evaluation Framework

Dimension	**Goal**	**Metrics**	**Methods**
Personalization & Adaptability	Ensure dynamic adjustment of tone, content, and recommendations without compromising accuracy or ethical integrity.	Task success rate per user segment, engagement duration, repeated usage frequency, personalization score	Controlled user studies combined with real-world interaction logs; adaptive scenario testing where user preferences evolve during the evaluation period.
Information Accuracy	Minimize misinformation and hallucinations while maintaining domain-specific accuracy.	Semantic similarity (cosine similarity, BERTScore), fact verification accuracy, correction rate post-feedback, cross-model consistency score.	Benchmark question sets with verified ground truth, multi-model consensus checks, real-time fact-checking modules.
Response Adequacy & Speed	Balance rapid response generation with contextual completeness and fluency.	Answerability score (−1 to +1), task completion time, latency to first response, BLEU/ROUGE scores for linguistic quality, distinct-n for lexical diversity.	Scripted and live user sessions measuring both efficiency and conversational coherence; stress testing under high query volume.
Bias & Ethical Considerations	Identify and mitigate unfair biases, protect user privacy, and maintain transparent, accountable system behavior.	Bias detection rate, privacy compliance score, transparency index, ethical category coverage.	Structured bias prompts, cross-cultural scenario testing, independent ethical audits, explainability reports for decision-making processes.

V. CONCLUSIONS

This study focused on the comparative analysis of AI chatbot evaluation frameworks. The findings confirm that a multidimensional approach is essential for evaluating this

technology, including factors like information accuracy, adaptability, ethical considerations, and user experience. In particular, key elements are the personalization and adaptability of these systems, as chatbots that can recognize and respond to user preferences and needs, significantly enhance usability. Similarly, the assessment of chatbots' accuracy and reliability is important for their acceptance, since these factors influence user trust and perceived usefulness. Moreover, ethical concerns, including bias detection and privacy protection, have to be evaluated in order to ensure fairness and transparency. As such, the development of an evaluation model that combines ethical techniques, functional metrics, human-in-the-loop methods, and user experience analysis appears to be the most effective approach.

Future steps include the systematic evaluation of the proposed evaluation framework in order to ensure that its structure and components effectively address the multidimensional nature of chatbot performance. In particular, future work will define specific, measurable metrics and appropriate evaluation methods tailored to each evaluation dimension. Moreover, the framework will be applied across various domains to test its adaptability and identify domain-specific challenges.

REFERENCES

[1] Casas, J., Tricot, M. O., Abou Khaled, O., Mugellini, E., & Cudré-Mauroux, P. (2020, October). Trends & methods in chatbot evaluation. In Companion Publication of the 2020 International Conference on Multimodal Interaction (pp. 280-286).

[2] Dodda, S. B., Maruthi, S., Yellu, R. R., Thuniki, P., & Reddy, S. R. B. (2021). Conversational AI-Chatbot Architectures and Evaluation: Analyzing architectures and evaluation methods for conversational AI systems, including chatbots, virtual assistants, and dialogue systems. Australian Journal of Machine Learning Research & Applications, 1(1), 13-20.

[3] Kim, H., Jung, S., & Ryu, G. (2020). A study on the restaurant recommendation service app based on AI chatbot using personalization information. International Journal of Advanced Culture Technology, 8(4), 263-270.

[4] Kuznetsova, E., Makhortykh, M., Vziatysheva, V., Stolze, M., Baghumyan, A., & Urman, A. (2025). In generative AI we trust: can chatbots effectively verify political information?. Journal of Computational Social Science, 8(1), 15.

[5] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., ... & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digital Medicine, 7(1), 258.

[6] Krouska, A., Troussas, C., Voyiatzis, I., Mylonas, P., & Sgouropoulou, C. (2024, November). ChatGPT-based recommendations for personalized content creation and instructional design with a tailored prompt generator. In 2024 2nd International Conference on Foundation and Large Language Models (FLLM) (pp. 295-299). IEEE.

[7] Gupta, P., Rajasekar, A. A., Patel, A., Kulkarni, M., Sunell, A., Kim, K., ... & Trivedi, A. (2022, December). Answerability: A custom metric for evaluating chatbot performance. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM) (pp. 316-325).

[8] Atkins, S., Badrie, I., & van Otterloo, S. (2021). Applying Ethical AI Frameworks in practice: Evaluating conversational AI chatbot solutions, Computers and Society Research Journal, 1. Retrieved May, 24, 2022.

[9] Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019, March). A survey on evaluation methods for chatbots. In Proceedings of the 2019 7th International Conference on Information and Education Technology (pp. 111-119).

[10] Peras, D. (2018). Chatbot evaluation metrics. Economic and Social Development: Book of Proceedings, 89-97.

[11] Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L., & Callison-Burch, C. (2019, June). Chateval: A tool for chatbot evaluation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (pp. 60-65).

[12] Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2023). The power of personalization: A systematic review of personality-adaptive chatbots. SN Computer Science, 4(5), 661.

[13] Izadi, S., & Forouzanfar, M. (2024). Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots. AI, 5(2), 803-841.

[14] Ghandour, A., Woodford, B. J., & Abusaimeh, H. (2024). Ethical Considerations in the Use of ChatGPT: An Exploration Through the Lens of Five Moral Dimensions. IEEE Access.

[15] Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. International Journal of Social Research Methodology, 8(1), 19–32. https://doi.org/10.1080/1364557032000119616

[16] Putri, F. P., Meidia, H., & Gunawan, D. (2019, December). Designing intelligent personalized chatbot for hotel services. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (pp. 468-472).

[17] Kadir, R. A., Zulkifli, M. F., Tiun, S. B., Lakulu, M. M., Jusoh, S., & Faudzi, A. F. A. (2023, September). EduChat: AI-powered chatbot with personalized engagement for online learning. In Intelligent Systems Conference (pp. 589-597). Cham: Springer Nature Switzerland.

[18] Iyenghar, P., Hu, Y., Kieviet, M., Pulvermueller, E., & Wuebbelmann, J. (2022, October). AI-Based Assistant for Determining the Required Performance Level for a Safety Function. In IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society (pp. 1-6). IEEE.

[19] Seidel, P., & Späthe, S. (2024, May). Development and Validation of AI-Driven NLP Algorithms for Chatbots in Requirement Engineering. In International Conference on Innovations for Community Services (pp. 132-149). Cham: Springer Nature Switzerland.

[20] Sharma, D., Reddy, N., Gupta, P., & Sharma, R. (2022). Enhancing Customer Experience Personalization through AI: Leveraging Collaborative Filtering, Neural Networks, and Natural Language Processing. Journal of AI ML Research, 11(7).

[21] Ahamed, B., Kareem, F., & Mohamed, M. Y. N. (2025, February). Innovative Approaches in Predictive Analysis and Personalized Online Shopping Recommendations with AI Powered-Chat. In 2025 International Conference for Artificial Intelligence, Applications, Innovation and Ethics (AI2E) (pp. 1-6). IEEE.

[22] Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. menu-based interfaces: An empirical study. Computers in Human Behavior, 128, 107093.

[23] Rani, Y. A., Balaram, A., Sirisha, M. R., Nabi, S. A., Renuka, P., & Kiran, A. (2024, April). AI Enhanced Customer Service Chatbot. In 2024 International Conference on Science Technology Engineering and Management (ICSTEM) (pp. 1-5). IEEE.

[24] Ma, Z., Dou, Z., Zhu, Y., Zhong, H., & Wen, J. R. (2021, July). One chatbot per person: Creating personalized chatbots based on implicit user profiles. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval (pp. 555-564).

[25] Chinedu-Eneh, P., & Nguyen, T. T. (2024, January). Elephant: LLM System for Accurate Recantations. In 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0191-0197). IEEE.

[26] Nambiar, J. P., & Sreedevi, A. G. (2023, November). Orchestrating Consensus Strategies to Counter AI Hallucination in Generative Chatbots. In 2023 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM) (pp. 148-152). IEEE.

[27] Williamson, S. M., & Prybutok, V. (2024). The era of artificial intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation. Information, 15(6), 299.

[28] Jacob, C., Kerrigan, P., & Bastos, M. (2025). The chat-chamber effect: Trusting the AI hallucination. Big Data & Society, 12(1), 20539517241306345.

[29] Gao, J., Agarwal, R., & Garsole, P. (2025). AI Testing for Intelligent Chatbots—A Case Study. Software, 4(2), 12.

[30] Chen, J. S., Le, T. T. Y., & Florence, D. (2021). Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing. International Journal of Retail & Distribution Management, 49(11), 1512-1531.

[31] Vijayaraghavan, V., & Cooper, J. B. (2020). Algorithm inspection for chatbot performance evaluation. Procedia Computer Science, 171, 2267-2274.

[32] Asiedu, E., Boakye, A. N., Malcalm, E., & Majeed, M. (2024, April). AI-Enabled Chatbot Integration on Business Process and Its Effect on Service Performance. In International Conference on Advances in Information Communication Technology & Computing (pp. 175-189). Singapore: Springer Nature Singapore.

[33] Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. Sustainability, 15(7), 5614.

[34] Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and Ethical Considerations. AICS, 2563, 104-115.

[35] Ahmed, I., Liu, W., Roscoe, R. D., Reilley, E., & McNamara, D. S. (2025). Multifaceted Assessment of Responsible Use and Bias in Language Models for Education. Computers, 14(3), 100.

[36] O'Leary, D. E. (2024). Do large language models bias human evaluations?. IEEE Intelligent Systems, 39(4), 83-87.

[37] Beattie, H., Watkins, L., Robinson, W. H., Rubin, A., & Watkins, S. (2022, March). Measuring and mitigating bias in AI-chatbots. In 2022 IEEE International Conference on Assured Autonomy (ICAA) (pp. 117-123). IEEE.

[38] Mhatre, A. (2023, October). Detecting the presence of social bias in GPT-3.5 using association tests. In 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA) (pp. 1-6). IEEE.