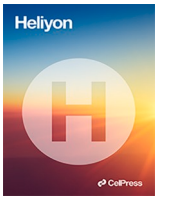
Heliyon 11 (2025) e42077

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Review article

Explainable AI chatbots towards XAI ChatGPT: A review

Attila Kovari [a,b,c,d,*]

[a] Institute of Digital Technology, Faculty of Computer Science, Eszterházy Károly Catholic University, Eszterhazy ter 1, Eger, 3300, Hungary
[b] Institute of Computer Engineering, University of Dunaújváros, Dunaújváros, Hungary, Tancsics M. 1/A, 2400, Dunaujvaros, Hungary
[c] Department of Informatics, GAMF Faculty of Engineering and Computer Science, John von Neumann University, Izsáki u. 10, 6000, Kecskemét, Hungary
[d] Institute of Electronics and Communication Systems, Kandó Kálmán Faculty of Electrical Engineering, Óbuda University, Bécsi street 96/B, 1034, Budapest, Hungary

ARTICLE INFO

Keywords:
Explainable AI (XAI)
ChatGPT
AI chatbots
Natural language processing (NLP)
Transparency
Controllable AI

ABSTRACT

Advances in artificial intelligence (AI) have had a major impact on natural language processing (NLP), even more so with the emergence of large-scale language models like ChatGPT. This paper aims to provide a critical review of explainable AI (XAI) methodologies for AI chatbots, with a particular focus on ChatGPT. Its main objectives are to investigate the applied methods that improve the explainability of AI chatbots, identify the challenges and limitations within them, and explore future research directions. Such goals emphasize the need for transparency and interpretability of AI systems to build trust with users and allow for accountability. While integrating such interdisciplinary methods, such as hybrid methods combining knowledge graphs with ChatGPT, enhancing explainability, they also highlight industry needs for explainability and user-centred design. This will be followed by a discussion of the balance between explainability and performance, then the role of human judgement, and finally the future of verifiable AI. These are the avenues through which insights can be used to guide the development of transparent, reliable and efficient AI chatbots.

1. Introduction

The evolution of artificial intelligence (AI) has significantly transformed various technological landscapes, particularly the field of natural language processing (NLP). One of the most important areas to develop this capability is AI-based chatbots, evolving from early systems such as ELIZA from the 1960s [1]. Early ELIZA used simple pattern matching and substitution methods to decipher conversation but did not really understand it and lacked contextual awareness.

Advances in machine learning and deep learning in recent decades have catalysed the research and development of more advanced chatbots and the number of articles is increasing (Fig. 1.). Furthermore, as AI technologies, such as ChatGPT, continue to develop, they no longer only affect technical domains; instead, they are reshaping more general areas of human life, including job loss and labour productivity [2], as well as philosophical debates about the future of humanity [3].

Modern chatbots need complex algorithms running on massive data sets to perform tasks such as automated customer service, personal virtual assistance, and content generation. One of the most prominent developments in recent years is the transformer architecture [4]. This architecture has formed the backbone of very many modern NLP models, including the OpenAI GPT-3 model [5].

* Institute of Digital Technology, Faculty of Computer Science, Eszterházy Károly Catholic University, Eszterhazy ter 1, Eger, 3300, Hungary.
E-mail addresses: kovari.attila@uni-eszterhazy.hu, kovari@uniduna.hu, kovari.attila@uni-obuda.hu, kovari.attila@nje.hu.

https://doi.org/10.1016/j.heliyon.2025.e42077
Received 1 August 2024; Received in revised form 1 January 2025; Accepted 16 January 2025
Available online 17 January 2025

The GPT-3 model uses 175 billion parameters to generate human-like text making it one of the most powerful language models to date. ChatGPT exploits this architecture to provide coherent and contextually relevant response sequences in a conversational environment. These developments represent a huge shift from rule-based to much more flexible and adaptive AI models that attempt to account for the nuances of understanding and language shaping. Nowadays the ChatGPT and GPT models are of great interest in academia (Fig. 2).

As AI systems such as ChatGPT become more and more embedded in everyday life, there is a growing concern about explainability and transparency. This is important not only for technical correctness but also for maintaining confidence in AI systems, especially when the applications are in sensitive employment areas, where the impact of AI on skills and well-being is a very important concern [6]. Work in XAI focuses on making AI decisions more meaningful, building trust and increasing accountability [7]. This becomes very important for chatbots, which users rely on to provide them with relevant information and sometimes support in their decision-making [8]. In the case where the answer is not clearly explained, the chatbot is perceived as untrustworthy - an aspect that laterally feeds into user distrust [9]. This is also extremely important for explaining and mitigating biases within AI systems [10]. AI models pick up biases in their training data with enough certainty to be able to propagate them into their decisions, which typically means unfair outcomes in sensitive applications such as hiring and credit approval [11]. Clear explanations therefore help users and developers to recognize and correct these biases to ensure fairness and ethicality when using AI. The application area, in particular e-commerce, healthcare, education and customer service, has been in use for quite some time with the involvement of AI-based chatbots such as ChatGPT. However, most of these existing AI models remain a "black box" in terms of transparency and accountability [12]. On these issues discussed above, XAI provides transparency to AI decision-making processes to build trust and facilitate wider adoption by users [13].

This review highlights the importance of explainability in AI chatbots and the ongoing efforts to make these systems more transparent and reliable. The paper is organized as follows: Section 2 comparatively assesses the critical differences in ChatGPT versus other chatbot technologies. Section 3 discusses the aspects related to explainability in AI systems, with a particular emphasis on chatbots. The methodology of the review process is presented within the context of Section 4. Section 5 summarizes the results of the reviewed studies, pointing out key trends and research gaps. Section 6 discusses the challenges and limitations identified, while Section 7 presents the conclusions and some possible future research directions.

2. ChatGPT vs chatbot

Chatbot is a general term for any computer program that can imitate human conversation. Chatbots range from rule-based systems to advanced deep learning models like GPT [14]. Chatbot training is not applicable, especially for industries, unless the data is customised. For rule-based chatbots, user interactions are compared with databases [15,16].

ChatGPT belongs to the family of GPT models developed by OpenAI, which entails deep learning algorithms that are exposed to huge amounts of textual data during training. GPT models facilitate the generation of human-like texts for advanced response or even other language-related tasks. Common applications for natural language understanding and generation are in scientific research, content writing, customer service and education [17]. While every ChatGPT is a chatbot, not every chatbot is ChatGPT. ChatGPT is a specific type of chatbot built on the capabilities of GPT models and possesses advanced language processing abilities.

3. Importance of explainability

XAI contributes to the interpretability of AI decisions, building trust and accountability [18]. This becomes even more critical for chatbots, which users rely on for information and decision-making. In the absence of such transparent explanations, chatbots can become untrustworthy, eroding user trust [19]. More and more researchers are exploring the potential and research of XAI (Fig. 3.).

Explicability is crucial in detecting and mitigating biases in AI systems [20]. Even for seemingly unbiased inputs of AI models, their outputs sometimes inadvertently reflect biases in the training data, which has a large impact on fairness [21]. Transparent explanations enable both users and developers to detect and mitigate these biases, thereby promoting fair and ethical AI use. Next-generation applications of XAI chatbots such as ChatGPT can be found in areas such as e-commerce, healthcare, education, and customer service. XAI can help address this issue by enabling users to understand the AI decision-making process and building user trust in AI for its wider adoption [22].

XAI ChatGPT is truly the epitome of a move towards XAI chatbots. Heavily influenced by the GPT architecture, ChatGPT has received widespread attention for its ability to generate human-like text. Despite this, however, the decision-making processes of GPT

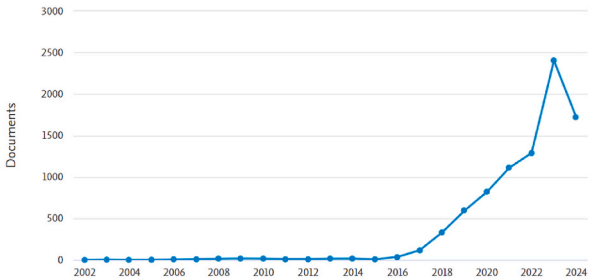


Fig. 1. Chatbot papers by year based on Scopus (Query: TITLE-ABS-KEY (chatbot)).

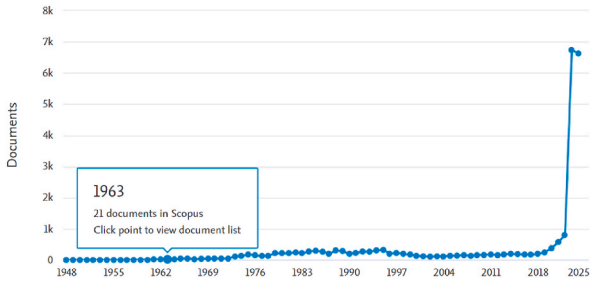


Fig. 2. ChatGPT papers by year based on Scopus TITLE-ABS-KEY (Query: TITLE-ABS-KEY (chatgpt OR gpt)).

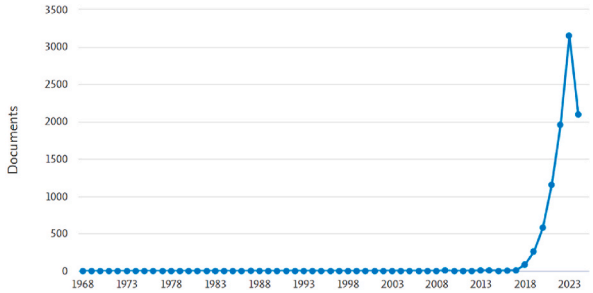


Fig. 3. XAI papers by year based on Scopus TITLE-ABS-KEY (Query: TITLE-ABS-KEY ("explainable AI" OR XAI)).

models are largely obscured for both the user and the developers [23]. This opacity therefore raises trust and accountability issues, especially in high-stakes environments such as healthcare and finance [24].

4. Scope and objectives

This review aims to provide a comprehensive overview of the current state of explainability in AI chatbots, with a particular focus on the ChatGPT model. The main objectives are.

(1) To examine the methodologies employed to enhance the explainability of AI chatbots.
(2) To identify the challenges and limitations associated with these methodologies.
(3) To explore potential future directions for research and development in XAI.

The increasing integration of XAI into chatbots highlights the need for a comprehensive review of current methodologies, applications and challenges with these systems. As this is a review article, the current literature on the topic of XAI-enabled chatbots and their cross-cutting applications needs to be organized and reviewed. The review will therefore also highlight how progress has been made so far and identify the gaps in the current results of this research with the intention of providing insights for the future development and implementation of XAI chatbots. The paper will present the broader implications of XAI within the current development and application of AI technologies, supporting the need for continued research in this area.

5. Methodology

5.1. Article search

This literature review is based on the Reporting Standards for Systematic Evidence Syntheses (ROSES) method [54]. The search for scientific articles was performed based on the Scopus database, among the most generally accepted Scopus and Web of Science databases, as it indexes a wider range of scientific literature, and its overlap with the Web of Science database is 99.1 % [25]. On July 12, 2024, a bibliographic search was conducted with the search query: TITLE-ABS-KEY ((chatgpt OR chatbot) AND (explainable OR xai)). This process yielded 126 articles.

5.2. Eligibility

The articles are examined to ascertain their adherence to the predetermined inclusion criteria. Specifically, the inclusion criteria for the current review encompassed studies that met the following conditions.

(1) studies related to Chatbot or ChatGPT;
(2) studies with a focus on explainable AI Chatbot or ChatGPT;
(3) as the topic is extremely hot and dynamic, articles selected from 2023 onwards;
(4) studies written in English.

A thorough screening of titles and abstracts from the initial list of 115 articles resulted in the exclusion of 31 conference summaries. Subsequently, upon comprehensive content review with the application of the inclusion and exclusion criteria, an additional 49 articles were excluded because these papers are related to general AI or not explainable AI focus. The culmination of this process yielded 29 articles eligible for the subsequent quality assessment step (see Fig. 4.). The document types of selected papers are shown in Fig. 5.

5.3. Summary of papers included in the review

Table 1 summarizes the details of the included papers, as follows: No; Publication year; Authors; Title. Based on a review of the articles one by one, other important factors related to the article have been identified and can also be found in Table 1: Focus Area; Methodology; Application Domain; Key Findings/Remarks; Explainable AI Implementation.

6. Results based on the literature review

The results organize the findings from a comprehensive literature review of XAI chatbots, focusing on the progress towards creating explainable versions of AI chatbots like ChatGPT. The aim will be to synthesize current efforts in research, methodologies and application domains to review the state of XAI in the context of chatbots. The authors of this paper identify common themes, significant achievements and emerging trends from a collection of very diverse studies contributing to transparent and trustworthy AI systems. In 2024, there were 12 publications among the peer-reviewed articles.

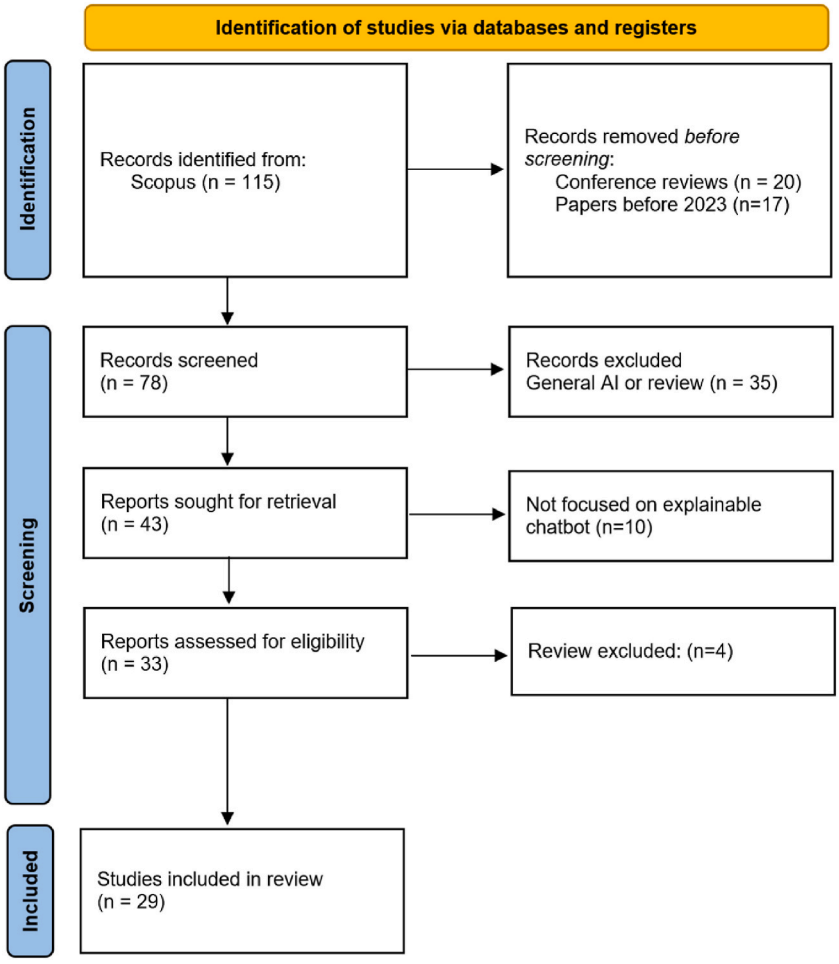


Fig. 4. Diagram of the search and selection process.

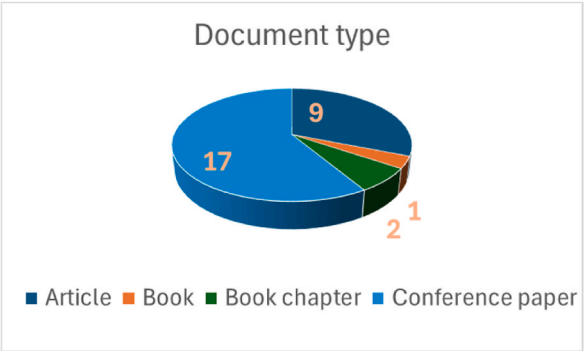


Fig. 5. Selected papers by type.

The distribution of the years of papers reviewed on XAI chatbots has been growing in interest over the years (Fig. 6.). While only one article was published in 2019, this number increased to two in 2020 and two in 2021. The real growth started in 2022 when six articles were published. In 2023, there was a significant jump, with 16 articles published, indicating increasing interest and research activity in the topic. For 2024, there are now 12 papers so far. This trend indicates that the field of XAI chatbots is dynamically developing and attracting more and more researchers.

The results of the reviewed studies address some key areas, the most important of which relate to the requirement for standardised evaluation of verifiable AI as a possible alternative approach, and the need for scalable and generalisable XAI techniques. The results highlight the critical role that human-AI collaboration can play in improving the quality of explanations and highlight the need for context-aware explainability techniques to be designed to suit the application domain [41]. These results show the state of the art in the field of XAI chatbots towards the development of AI systems that are both robust, transparent and user-friendly.

6.1. Integration of diverse methodologies enhances explainability

The results underscore the significant benefits of integrating diverse methodologies. This approach involves combining different techniques, such as knowledge graphs with ChatGPT [26] and utilizing generative adversarial networks (GANs) [29], to enhance the explainability of AI chatbots. The results indicate that hybrid approaches contribute to the development of more transparent and easily understandable AI systems to users, improving user trust and satisfaction [37,47].

Knowledge graphs are structured representations of knowledge designed to capture the relationships between concepts within a domain. The integration of these graphs with a very powerful language model, such as ChatGPT, allows the hybrid system to use structured knowledge to generate more contextually relevant and accurate explanations. For example, in applications such as planning and customer segmentation, domain-specific knowledge graphs help ChatGPT understand domain-related terminologies and the relationships between them to provide not only accurate but also contextually relevant explanations [18]. These methods also allow the chatbot to explain its reasoning processes in a way that best fits the user's understanding of the domain. The structured nature of the knowledge graphs can be easily used to break down complex explanations into simpler parts that are easier to digest and thus easy for the user to understand.

GANs consist of two neural networks: a generator and a discriminator, which work together to produce high-realism data patterns; they can be used in explainability to produce feasible alternative scenarios for users to understand AI decision-making. For example, in most domains, GANs can be used to generate counterfactual explanations by constructing hypothetical scenarios through which it would be discernible that different outcomes will occur for different inputs [29]. In this way, the user can gain insight into the AI system's behaviour and recognize the potential impact of different factors on their decisions. This means that by using GANs to generate such counterfactual explanations, it is ensured that they will not only be realistic but also relevant; thus, providing better transparency for the AI systems as a whole.

6.2. Sector-specific explainability needs

The findings reveal that different application domains have unique requirements for explainability [52]. These findings underscore the importance of tailoring techniques of explainability to sector-specific needs to make the relevance and effectiveness real. Explanatory techniques are important for developing AI systems that are context-aware and, thus, user-centred so that transparency can be achieved along with trustworthiness and effectiveness across a wide spectrum of domains.

The paper [30] discuss the concept of controllable AI as a method to ensure trustworthiness. They also demonstrate that explainability will be essential for the development of verifiable and reliable AI systems. If an AI system can provide understandable information about its decision-making process through explainable features, then trust and user satisfaction will increase. Specifically, such research shows that in order to be trusted and reliable by the user, explainability needs to be integrated into an AI system to derive insights. The researcher on e-commerce applications of XAI found that the explainability features built into an XAI chatbot dramatically improve users' trust in product recommendations [27]. This increase in confidence results in satisfaction as the user has more

trust in the suggestions coming from the chatbot [53]. This shows how explainability can directly impact the user experience in practical, commercial contexts.

For instance, in the field of mental health analytics, large-scale language models have been used to create meaningful analyses; this increases the relevance and power of explanations [28]. In the medical field, notably using interpretable AI techniques such as CureNet, domain-specific large-scale language models have been used to increase the accuracy and interpretability of diagnoses by increasing trust between healthcare professionals [35].

6.3. User trust and satisfaction

Clarity in answers builds trust in the AI system because users trust the guidance presented to them [52]. In general, this can lead to a better user experience with such systems. Explicability is one of the most important elements in building user trust in AI technologies. The paper [31] present the development of explainable agents with a framework that can return an understandable rationale for their actions at any time, thus increasing user trust.

Explainable conversational agents contribute significantly to improving user understanding and fostering trust in acceptance, addressing the black box problem of AI algorithms, and underlining the fact that the road to more user-friendly AI systems is transparency and rationality [46]. Integrating Socratic questioning methods into the advanced capability of ChatGPT that the proposal will design a far more active and multilevel pattern of dialogue structure that will meet the cognitive and emotional needs of older adults. WisCompanion, in furthering critical thinking, trust, and transparency in interactions using AI, can greatly enhance user satisfaction with emotional support to foster lifelong learning and further improve the quality of the overall interaction [48].

A combination of different explanation strategies depending on different user needs and greatly increases trust and satisfaction with AI-assisted decisions. The iSee platform, developed in the framework of a research project, implements this principle in its basic functions with the help of an ontology-guided reasoning system, which is further developed by user feedback in the form of a conversational chatbot [50]. This integration also helps support XAI best practices and effectively captures and manages certain annotation requirements, as demonstrated in a case study based on an X-ray fracture detection system. Additionally, LLM4V is a ChatGPT-based explainable approach to visualization recommendations increasing trust among users by making the decision-making process transparent [32]. For ChatGPT, explainability in the context of facial biometric data analysis is a key aspect in building increased trust in AI-driven recognition systems [40].

6.4. Challenges in balancing explainability and performance

This also raises the issue of the trade-offs between explainability and performance in the development of AI-enabled systems. Improved transparency bridges the gap in user trust and ensures ethics-based decision-making, but not necessarily at the expense of system efficiency and effectiveness. In the following, we review how to optimise precisely this balance in the development of highly efficient yet transparent AI systems.

Several studies have explored various methodologies to achieve this balance. The paper [37] proposed a framework of transparent predictive analytics that emphasizes explainability, which is introduced without compromising on the accuracy of the models. A combined blockchain and XAI techniques is applied to increase safety and transparency across applications and ensure that performance metrics remain robust [43]. The paper [33] analysed researcher sentiments about ChatGPT and the results showed the need for AI models that are more transparent and without sacrificing performance.

The research [39] provided an overview of the application of AI in biomedical science, where there is a strong need for explainable models in high-stakes environments (as performance cannot afford to compromise). In the legal domain, they proposed combining blockchain with XAI for enhanced security, and thus to become more transparent, ensuring robust performance measures [43].

It was also found that large-scale language models, complemented with explainability, can increase transparency and collaboration in digital advertising [47]. Another study looked at the integration of AI into decision models for explainability, in a range of domains such as health, finance and law [23].

6.5. Human evaluations as a benchmark

Human evaluation is essential to establish a baseline for the effectiveness of explanatory methods [41]. Such an evaluation alone will provide meaningful explanations that improve user interaction with AI systems. The paper [34] studied the reasoning of AI, they use human judgement to probe whether AI produces logical and understandable explanations to elicit answers, especially for question-answering systems.

AI diagnostics is supported by the CureNet tool, which aims to make answers more explainable in healthcare regardless of the path to the result [35]. Human assessment will be paramount in this regard to ensure that the explanatory accuracy provided by such AI is not only correct but also comprehensible to medical practitioners. The application of ChatGPT's code interpreter in the medical field show improvements in the explanations provided by this system with human feedback, which would provide better support for the medical decision-making process [36]. This demonstrates the role that human judgement can play in enhancing the clarity and relevance of AI explanations, building user trust and encouraging greater engagement.

By focusing on human evaluations, these explanations make sense in terms of user interaction with the chatbot. A qualitative framework for applications of XAI and machine learning is used to further emphasize the need for human evaluations to demonstrate the effectiveness of various XAI methods [49].

6.6. Applications in various domains

The reviewed literature covers a wide range of application domains, each with unique requirements and methodologies for implementing XAI and this section summarizes these applications.

In the field of IT, explainability has been shown to be key to building increased trust in AI-driven recognition systems [40]. Focusing on an overview of the trends and challenges of artificial intelligence applications in biotechnology, authors highlight the importance of the requirement of explainable models in high-stakes situations [39]. This paper discusses the trends and challenges in applying AI for biotechnological applications, emphasizing the need for explainable models in high-stakes environments. The limitations of current transparency and explainability efforts in addressing the diverse needs of different user groups and proposes a user-centric framework that emphasizes practical guidelines on how to effectively and responsibly use LLMs [45]. Voice-based conversational recommendation systems are an improvement over text-based recommendations and offer a fairly intuitive, inclusive, and natural user interaction [51]. The research focuses on the transparency and comprehensibility of user interactions, which not only provide a natural and accessible user experience, but are also self-explanatory, so users can easily understand the logic and reasons behind the recommendations.

In e-commerce, XAI chatbots have shown significant improvements in increasing user trust in product recommendations, and thus improving customer satisfaction and loyalty and enable transparency and effectiveness of advertising strategies [27]. In the legal field, it is proposed to create an XAI-based justice system using blockchain to increase the transparency and certainty of legal decision-making processes [43]. In the legal field, it is proposed to create an XAI-based justice system using blockchain to increase the transparency and certainty of legal decision-making processes [43]. The drivers of users' purchase decisions and willingness to pay were examined in connection with an XAI-based system, emphasizing that the interactivity and reliability of the system represent significant value for the target group in addition to transparency and explanatory nature [44].

The exploration of controllable AI represents a critical future direction in XAI research, emphasizing the development of AI systems capable of adapting their behaviour and explanations in real-time based on user feedback [30]. This dynamic adaptability will make AI systems always open, explainable, responsive and effectively applicable in different domains. Within quality assurance tasks, they focus on LLMs' use of common sense in reasoning and explainability, thus illustrating how LLMs can provide human-like explanations and make AI-based decisions more meaningful [34].

Code-based queries help explain the use of ChatGPT in supporting medical decision-making, showing how AI arrives at reasoning, thereby making healthcare more transparent [36]. In the field of education, a construct for supporting at-risk students is analysed by integrating XAI and prescriptive analytics, demonstrating the use of XAI in learning analytics [37]. They also analysed the effects of knowledge-driven AI technologies to benefit human life by using neurosymbolic AI and XAI to improve understandability and usability to improve multiple perspectives of human life [38].

In social media sentiment analysis, graph embedding approaches can be applied with machine learning algorithms to develop sentiment analysis models that can be explained to users, recognizing predicted causes of sentiment and increasing trust and satisfaction [42].

7. Discussion

The increasing complexity of AI systems and their integration into everyday life applications underlines the urgent need for XAI in technologies used by chatbots such as ChatGPT. This further highlights the importance for users of the need for more informed decision-making and recommendations as AI chatbots are being applied in many other areas. The aim of this review has been to present methods for improving the interpretability of AI chatbots in general, to identify challenges and limitations associated with these methods, and to explore possible future research and development directions in this area.

Some important lessons from this review are some of the new innovative methodologies being implemented to increase the explainability of AI chatbots. Some of these hybrid approaches show great promise, such as knowledge graphs integrated with ChatGPT. Knowledge graphs help to further organize any information in a truly accessible and meaningful way, allowing ChatGPT to offer contextually relevant explanations [26]. This not only increases the transparency of the decision-making process of the chatbot but also tailors the explanations to the user's knowledge, thus increasing the overall satisfaction of the users.

Despite the progress made with XAI chatbots, many challenges and limitations remain in their development. A key challenge is the trade-off between explainability and performance. If transparency is required to gain the trust of users, this cannot be at the expense of the efficiency and effectiveness of AI systems. In other words, ensuring that explainable models perform on par with their opaque counterparts is another huge hurdle that researchers continue to address. In addition, there is a need for evaluation metrics that provide built-in features for assessing the quality and effectiveness of explanations in AI systems. In the absence of a good explanation, comparing and complementarily improving XAI methods is indeed very difficult. The development of robust, domain-independent metrics will be important in the development of this field.

The application areas of explainability differ in their specific requirements and therefore require application-specific approaches. For example, explainability is relevant for diagnostic tools used in the healthcare industry, where healthcare professionals need to trust and understand recommendations driven by AI [35]. In contrast, e-commerce chatbots provide explainability that enables customers to understand different product recommendations, which increases their trust and leads to higher user satisfaction [27]. Such a diversity of needs highlights the need to develop context-aware interpretability solutions that understand the challenges of each domain.

XAI methodologies can be tailored to the needs of various applications, considering the specific requirements and challenges of each application. In the field of education, XAI can be designed to provide pedagogically structured, step-by-step explanations, where

Table 1
Summary of reviewed studies.

No	Year	Authors	Title	Focus Area	Methodology	Application Domain	Key Findings/ Remarks	Explainable AI Implementation
1 [26]	2024	Hu X., Liu A., Dai Y.	Combining ChatGPT and knowledge graph for explainable machine learning-driven design: a case study	Explainable AI in Design	A hybrid approach combining ChatGPT and knowledge graph	Design and customer segmentation	The hybrid method improves explainability by generating contextual and meaningful explanations in the design context.	Utilizes domain-specific knowledge graphs and ChatGPT for posthoc explainability
2 [27]	2024	Thapliyal K., Thapliyal M.	Chatbot-XAI? The New Age Artificial Intelligence Communication Tool for E-Commerce	Explainable AI in E-commerce	Natural language processing (NLP) and machine learning (ML) algorithms	E-commerce	XAI chatbots can explain their recommendations and improve customer trust and transparency in online shopping.	Explainability features are built into chatbots to clarify decision-making processes
3 [28]	2023	Yang K., Ji S., Zhang T., Xie Q., Kuang Z., Ananiadou S.	Towards Interpretable Mental Health Analysis with Large Language Models	Mental Health Analysis	Large language models (LLMs) with prompting strategies	Mental health	ChatGPT shows strong learning ability and generates human-like explanations, improving mental health analysis.	Uses human evaluations and prompting strategies to generate explanations for mental health decisions
4 [29]	2024	Del Ser J. et al.	On generating trustworthy counterfactual explanations	Counterfactual Explanations	Generative adversarial network with multiobjective optimization	Various domains	The framework balances plausibility, intensity of changes, and adversarial power to generate trustworthy counterfactual explanations.	Combines generative adversarial networks with multiobjective optimization
5 [30]	2023	Kieseberg P. et al.	Controllable AI - An Alternative to Trustworthiness in Complex AI Systems?	Controllable AI	Various techniques for implementing controllable AI	Various domains	Discusses the concept of controllable AI as an alternative to trustworthy AI and explores techniques for achieving it.	Focuses on implementing controllable AI techniques
6 [31]	2024	Rodriguez S., Thangarajah J.	Explainable Agents (XAg) by Design	Explainable Agents	Design patterns for explainable agents	Various domains	Proposes a framework for developing agents that are explainable by design using design patterns.	Advocates for explainable agents developed using design patterns
7 [32]	2023	Wang L., Zhang S., Wang Y., Lim E.-P., Wang Y.	LLM4Vis: Explainable Visualization Recommendation using ChatGPT	Visualization Recommendation	ChatGPT-based prompting approach	Data visualization	Proposes LLM4Vis, a ChatGPT-based approach for visualization recommendation with human-like explanations.	Uses ChatGPT to generate explanations for visualization recommendations
8 [33]	2024	Lakhanpal S., Gupta A., Agrawal R.	Leveraging Explainable AI to Analyze Researchers' Aspect-Based Sentiment About ChatGPT	Sentiment Analysis	Explainable AI for aspect-based sentiment analysis	Research analysis	Analyzes researcher sentiments about ChatGPT using aspect-based sentiment analysis with explainable AI techniques.	Employs explainable AI techniques for sentiment analysis in research
9 [34]	2024	Krause S., Stolzenburg F.	Commonsense Reasoning and Explainable Artificial	Commonsense Reasoning	Large language models with commonsense reasoning	Question answering (QA)	Investigates ChatGPT's ability to perform commonsense	Uses large language models for common

(*continued on next page*)

Table 1 (*continued*)

No	Year	Authors	Title	Focus Area	Methodology	Application Domain	Key Findings/ Remarks	Explainable AI Implementation
			Intelligence Using Large Language Models				reasoning and provide explanations for QA tasks.	reasoning and explainability
10 [35]	2023	Khambampati S., Dondapati S., Kattamuri T. V., Pathinarupothi R.K.	CureNet: Improving Explainability of AI Diagnosis Using Custom Large Language Models	Medical Diagnosis	Custom large language models with XAI techniques	Healthcare	Introduces CureNet, a tool integrating DL models with XAI techniques for accurate and explainable cardiac diagnosis.	Combines DL models with XAI techniques for medical diagnosis
11 [36]	2023	Kitamura K., Irvan M., Shigetomi Yamaguchi R.	XAI for Medicine by ChatGPT Code interpreter	Medical Decision-making	Code Base Prompt (CBP) using ChatGPT Code interpreter	Medical field	Proposes CBP to make ChatGPT's medical decision-making process explainable using Python code execution.	Uses ChatGPT Code interpreter for medical decision explainability
12 [37]	2023	Susnjak T.	Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT	Learning Analytics	Transparent machine learning with prescriptive analytics	Education	Proposes a framework for transparent predictive modelling and prescriptive analytics to support at-risk students.	Integrates XAI with prescriptive analytics for educational support
13 [38]	2023	Ukil A., Jara A. J., Gama J., Marin L.	Knowledge-driven Analytics and Systems Impacting Human Quality of Life - Neurosymbolic AI, Explainable AI and Beyond	Human Quality of Life	Neurosymbolic AI and XAI	Various domains	Discusses the impact of AI on human life and presents research findings on knowledge-driven AI technologies.	Explores the impact of neurosymbolic AI and XAI on human life
14 [39]	2023	Holzinger A. et al.	AI for life: Trends in artificial intelligence for biotechnology	AI for Biotechnology	Machine learning, NLP, decision support	Biotechnology	Provides an overview of AI applications in biotechnology and discusses open research issues.	Explores methodological aspects of XAI in biotechnology
15 [40]	2024	DeAndres-Tame I. et al.	How Good Is ChatGPT at Face Biometrics? A First Look into Recognition, Soft Biometrics, and Explainability	Face Biometrics	GPT-4 multimodal LLM	Biometrics	Analyzes ChatGPT's performance in face verification and explainability of results.	Uses GPT-4 for explainability in face biometrics
16 [41]	2024	Sung Y.-H., Chien S.-Y., Yu F.	Preliminary Validation of Explainable AI Interfaces across Heuristics and Information Transparency	Human-Chatbot Interaction	SAT model for transparency	E-commerce	Investigates heuristic mechanisms and transparency in AI recommendations.	Uses SAT model for transparency in human-chatbot interaction
17 [42]	2024	Anoop V.S., Subin Krishna C., Govindarajan U.H.	Graph embedding approaches for social media sentiment analysis with model explanation	Social Media Sentiment Analysis	Graph neural network embeddings with machine learning algorithms	Social media	The approach uses SHAP for explaining sentiment analysis models and publishes labeled datasets for further research.	Uses SHAP to explain the output of machine learning models
18 [43]	2023	Demertzis K., Rantos K., Magafas L.,	A Secure and Privacy-Preserving	Justice System	Blockchain and privacy techniques	Legal domain	Proposes a framework combining	Uses blockchain and XAI to enhance security

(*continued on next page*)

Table 1 (*continued*)

No	Year	Authors	Title	Focus Area	Methodology	Application Domain	Key Findings/ Remarks	Explainable AI Implementation
		Skianis C., Iliadis L.	Blockchain-Based XAI-Justice System		combined with XAI		blockchain and XAI for a secure, transparent justice system.	and transparency in legal decisions
19 [44]	2023	Berger P., von Garrel J.	How to design a value-based Chatbot for the manufacturing industry: An empirical study of an internal assistance for employees	Manufacturing Industry	Choice-based conjoint analysis (CBC)	Business organizations	Analyzes user drivers and willingness to pay for value-based, explainable chatbots in manufacturing.	Focuses on transparency and interactivity in chatbot design for manufacturing
20 [45]	2024	Barman K.G., Wood N., Pawlowski P.	Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use	User Guidelines for LLMs	Framework for user-centric guidelines	Various domains	Proposes a framework for contextualized user guidelines to manage risks of LLM misuse.	Emphasizes education and contextualized guidelines for responsible LLM use
21 [46]	2024	Joshi R., Graefe J., Kraus M., Bengler K.	Exploring the Impact of Explainability on Trust and Acceptance of Conversational Agents – A Wizard of Oz Study	Trust in Conversational Agents	Wizard of Oz study with between-subjects design	Vacation planning	Finds that high explainability increases trust and acceptance of conversational agents.	Uses a Wizard of Oz study to assess the impact of explainability on user trust
22 [47]	2023	Yang Q., Ongpin M., Nikolenko S., Huang A., Farseev A.	Against Opacity: Explainable AI and Large Language Models for Effective Digital Advertising	Digital Advertising	Combining LLMs with explainability features	Advertising	Proposes SODA system to enhance transparency and collaboration in digital advertising using LLMs.	Uses LLMs and explainability features to improve digital advertising strategies
23 [10]	2024	Pafa M., Larson K., Hancock M.	Unraveling the Dilemma of AI Errors: Exploring the Effectiveness of Human and Machine Explanations for Large Language Models	AI Error Explanation	Comparison of human and machine-generated explanations	Question answering	Finds that human saliency maps are more helpful than machine explanations, but trust decreases with wrong AI predictions.	Compares human and machine-generated explanations for AI errors
24 [48]	2024	Etori N.A., Gini M.	WisCompanion: Integrating the Socratic Method with ChatGPT-Based AI for Enhanced Explainability in Emotional Support for Older Adults	Emotional Support for Older Adults	Socratic method integrated with ChatGPT	Emotional support	Developed WisCompanion to provide tailored emotional support using Socratic questioning and ChatGPT.	Combines Socratic method with ChatGPT for enhanced emotional support
25 [49]	2023	Ramachandran M.	Quality framework for explainable artificial intelligence (XAI) and machine learning applications	XAI Quality Framework	Software engineering framework	Various domains	Proposes SEF-AI and ML framework to standardize XAI application development and improve quality.	Develops a software engineering framework for high-quality XAI applications
26 [50]	2023	Wijekoon A. et al.	CBR Driven Interactive Explainable AI	Interactive Explainable AI	Case-based reasoning (CBR) paradigm	Radiograph fracture detection	Introduces iSee, a platform for interactive explanation strategies using	Uses CBR for interactive and reusable explanation strategies

(*continued on next page*)

Table 1 (*continued*)

No	Year	Authors	Title	Focus Area	Methodology	Application Domain	Key Findings/ Remarks	Explainable AI Implementation
							CBR in radiograph detection.	
27 [51]	2023	Qu X., Liu H., Sun Z., Yin X., Ong Y.S., Lu L., Ma Z.	Towards Building Voice-based Conversational Recommender Systems: Datasets, Potential Solutions, and Prospects	Voice-based Recommender Systems	ChatGPT-driven prompts and text-to-speech model	E-commerce and movie domains	Creates datasets and explores solutions for voice-based conversational recommender systems.	Uses ChatGPT for generating voice-based recommendations
28 [52]	2023	Banafa A.	Transformative AI: Responsible, Transparent, and Trustworthy AI Systems	Responsible AI	Overview of trends and challenges	Various domains	Discusses trends, challenges, and opportunities in responsible AI systems, emphasizing transparency and trustworthiness.	Provides an overview of responsible AI practices and future directions
29 [23]	2023	Goossens A., Vanthienen J.	Integrating GPT-Technologies with Decision Models for Explainability	Decision Models	Integration of GPT-3 with decision models	Healthcare, finance, law	Explores the use of GPT-3 to identify appropriate reasoning and explanation scenarios for decision models.	Combines GPT-3 with decision models for enhanced explainability

everything is explained step-by-step as it would be taught in class. Providing short and clear explanations with XAI that are perfectly suited to the customer's question makes it easier to understand and promotes a timely response in this area. XAI methodologies can be designed to provide explanations sensitive enough to embed users' emotional states through support-oriented responses that are non-judgmental to gain trust and rapport. Distinct domains would require varying degrees of clarity, empathy, depth, and speed with which to display the explanation, thereby supporting the customization of XAI approaches to optimally address and serve the relevant audience.

The positive long-term results of using XAI chatbots in terms of user trust and satisfaction are overall positive; however, such behaviour develops under the influence of many factors. Above all, explainable AI chatbots develop trust due to the transparency of the reasons for their responses, thus ensuring a high satisfaction rate for the user in the early stages. This trust is further strengthened, as the chatbot continuously provides users with clear and comprehensible explanations over time, which strengthens and maintains, or even increases, satisfaction. However, over time this may lead to higher expectations due to increased familiarity with AI, so more advanced or contextually relevant explanations are needed to maintain the same level of satisfaction and trust. For this, continuous feedback loops and the iterative development of the XAI model can be of great help in gaining and increasing trust in the long term, so the development of the chatbot is guaranteed in addition to changing user needs. Therefore, the use of AI chatbots explains the improvement in user trust and satisfaction in the long term; these effects are dynamic and require continuous adaptation to user expectations and needs.

The future of XAI research lies in the development of controllable AI systems that allow for behaviour updates and real-time explanations as the user provides additional feedback. This would not only be responsive and transparent but would also guarantee

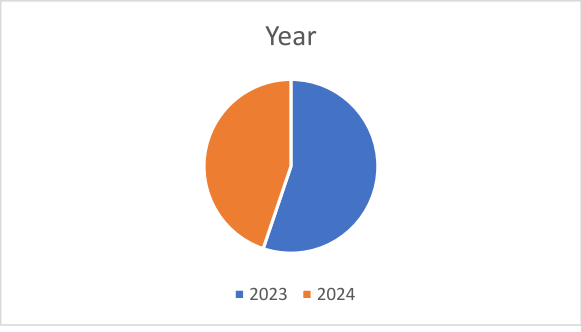


Fig. 6. Number of reviewed papers by year.

system efficiency and increase user confidence in several areas. Real-time adaptability, as seen in applications using large language models for commonsense reasoning [34], represents a promising direction for future research. Furthermore, the prospects of integrating XAI into new technologies such as blockchain point to opportunities to improve security and transparency to achieve safe and XAI-driven decisions, as two major problems that are desirable to address together in current AI applications.

The findings of this review have significant implications for both the development of AI systems and the use of AI in industries. As indicated throughout this paper, explainability in AI is not a technological issue that deals with technical correctness but an ethical, social, and regulatory one. First, transparency of AI's decision-making process is highly desired in these industries where AI is becoming an integral part, such as healthcare, finance, and e-commerce. It is in these domains, where decisions might implicate personal well-being, financial stability, or customer trust, that XAI offers a pathway toward increased accountability and fairness.

Secondly, the integration of explainability in AI systems allows for reshaping user interaction into increased trust and, thus, adoption of AI technologies. One underlying motivation is that as AI becomes more pervasive in personal and professional settings, so is the growing need for users to be presented intelligible, easy-to-understand reasoning related to AI-driven decisions. The openness that XAI brings enhances acceptance of AI in decision-making processes and, therefore, develops effective collaboration between human users and AI.

The role of XAI will further be called for in the mitigation of risks related to bias, discrimination, and those that involve unintended consequences as AI technologies become pervasive in high-stakes decisions, such as in legal judgments and autonomous systems. Our results indicate that further work on XAI methodologies is needed to keep their pace with the ever-growing complexity of AI models. Consequently, this may make the development of robust, sector-specific explainability standards and regulatory frameworks become priority issues among researchers and policymakers.

Finally, the prospect of XAI in the context of human-AI collaboration promises a world where algorithms will not only support but also educate and enable human users with meaningful insights into which action to take. This may shift the paradigm from AI as a tool to one in which AI is closer to an integral partner within decision-making processes, where it will have reshaping potential regarding how AI is made and perceived in society.

7.1. Challenges and limitations

Despite progress in XAI, there are still significant challenges. Most prominent is the trade-off between model complexity and interpretability. Models such as GPT-4 can achieve high performance, but their large number of parameters and complex internal representation make them very difficult to interpret. These models can be simplified towards interpretability but at the direct cost of a loss of accuracy and efficiency - which therefore poses another dilemma for AI system designers [22].

Another issue that is considered challenging is that most of the techniques within XAI are not very scalable in nature. Some of what has been discovered in explainability is challenging to scale for large, complex AI models. This is of more relevance for larger models such as ChatGPT, which need to be adapted to real-time interaction, with more robust and scalable approaches to generate meaningful explanations.

Another big issue is the user accessibility of the interpretability tools for non-experts. Most of the techniques previously developed for XAI are technical and require expertise to interpret them correctly. The key to increasing trust and widespread adoption of AI systems will be the development of intuitive interfaces and visualisations that enable end-users to understand AI decisions without deep technical expertise.

7.2. Future directions

The future of XAI in chatbots and broader AI applications is very promising and there are many research and development opportunities. On the one hand, XAI techniques need to be built in from the very beginning of the AI development lifecycle to become an integral part of AI systems. Another important requirement for making progress in this area is standardisation in the form of metrics and benchmarks that provide standardized and comparable metrics for transparency across models and applications.

This is particularly true for the emerging techniques of interactive explanation and dynamic context adjustment, which are described in the following sections. Interactive explanation allows users to query in real-time to gain insight into specific decisions to better understand model behaviour. A challenge that has been overlooked so far with respect to explainability is how to handle dynamic, multi-round interactions in chatbots. It will be important to ensure that explanations remain coherent and relevant in extended conversations.

Of course, ethical and regulatory issues are likely to be the main drivers behind XAI in the future. However, given the proliferation of AI systems, compliance with ethical standards and regulatory requirements for transparency and accountability will be very important. Finally, the development of frameworks and guidelines for the responsible use of AI will mitigate these risks, provided that the use of these technologies is done in a fair and transparent manner.

8. Conclusions

This review provided an overview of the current state of the art of explaining AI chatbots, such as ChatGPT. The main objectives in this context were: to identify the methods used to make AI chatbots more explainable, to identify the challenges and limitations of the methods used, and to identify possible future directions for further research and development of XAI.

The results show a range of methodologies, industry-specific needs and emerging trends in XAI chatbots. Hybrid approaches

leverage structured knowledge with new techniques to provide more meaningful and understandable explanatory models that bridge the gap between performance and interpretability in AI. User trust and interaction with clear, understandable explanations can be enhanced if AI chatbots are adopted. Attention-based models and explainable agents are just a few of the very many user-centric orientation principles within AI that make systems more transparent and indeed more trustworthy. The review underlines that finding the balance between explainability and performance is one of the key challenges that will need to be addressed rigorously in the future in order to ensure the efficiency and interpretability of AI systems.

Following these challenges and exploiting these opportunities, researchers and practitioners are pushing the boundaries of XAI to advance the development of AI systems that can be both efficient and yet transparent, given the need for trustworthy systems. This will be an initiative to realise the full potential of AI technologies in a way that meets user expectations and ethical standards.

Data availability

No data was used for the research described in the article.

Ethics statement

This article is a review of previously published research and does not involve any original experimental studies with human or animal subjects. Therefore, no ethical approval was required for this work.

Funding statement

The author received no financial support for the research, authorship, and/or publication of this article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Weizenbaum, ELIZA—a computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1) (1966) 36–45, https://doi.org/10.1145/365153.365168.
[2] G. Lazaroiu, E. Rogalska, How generative artificial intelligence technologies shape partial job displacement and labor productivity growth, Oeconomia Copernicana 14 (3) (2023) 703–706, https://doi.org/10.24136/oc.2023.020.
[3] M. Peters, et al., AI and the future of humanity: ChatGPT-4, philosophy and education–Critical responses, Educ. Philos. Theor. 56 (9) (2024) 828–862, https://doi.org/10.1080/00131857.2023.2213437.
[4] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008, in: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[5] T. Brown, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901. https://dl.acm.org/doi/abs/10.5555/3495724.3495883.
[6] R.E. Cramarenco, M.I. Burcă-Voicu, D.C. Dabija, The impact of artificial intelligence (AI) on employees' skills and well-being in global labor markets: a systematic review, Oeconomia Copernicana 14 (3) (2023) 731–767, https://doi.org/10.24136/oc.2023.022.
[7] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, A.I. Explainable, A brief survey on history, research areas, approaches and challenges, Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8 (2019) 563–574, https://doi.org/10.1007/978-3-030-32236-6_51.
[8] Y. Yamamoto, Suggestive answers strategy in human-chatbot interaction: a route to engaged critical decision making, Front. Psychol. 15 (2024) 1382234, https://doi.org/10.3389/fpsyg.2024.1382234.
[9] S. Pal, M. Bhattacharya, S.S. Lee, C. Chakraborty, A domain-specific next-generation large language model (LLM) or ChatGPT is required for biomedical engineering and research, Ann. Biomed. Eng. 52 (3) (2024) 451–454, https://doi.org/10.1007/s10439-023-03306-x.
[10] M. Pafla, K. Larson, M. Hancock, Unraveling the dilemma of AI errors: exploring the effectiveness of human and machine explanations for large language models, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–20, https://doi.org/10.1145/3613904.3642934.
[11] P.M. VanNostrand, D.M. Hofmann, L. Ma, E.A. Rundensteiner, Actionable recourse for automated decisions: examining the effects of counterfactual explanation type and presentation on lay user understanding, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1682–1700, https://doi.org/10.1145/3630106.3658997.
[12] J.R.R. Kumar, A. Kalnawat, A.M. Pawar, V.D. Jadhav, P. Srilatha, V. Khetani, Transparency in algorithmic decision-making: interpretable models for ethical accountability, in: E3S Web of Conferences, 491, EDP Sciences, 2024 02041, https://doi.org/10.1051/e3sconf/202449102041.
[13] M. Nagahisarchoghaei, N. Nur, L. Cummins, N. Nur, M.M. Karimi, S. Nandanwar, S. Rahimi, An empirical survey on explainable AI technologies: recent trends, use-cases, and categories from technical and application perspectives, Electronics 12 (5) (2023) 1092, https://doi.org/10.3390/electronics12051092.
[14] A. Benedek Szabo, J. Katona, A machine learning approach for skin lesion classification on iOS: implementing and optimizing a convolutional transfer learning model with Create ML, Int. J. Comput. Appl. (2024) 1–20, https://doi.org/10.1080/1206212X.2024.2380664.
[15] E. Adamopoulou, L. Moussiades, An overview of chatbot technology, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, 2020, pp. 373–383, https://doi.org/10.1007/978-3-030-49186-4_31.
[16] Gy Molnar, Z. Szuts, The role of chatbots in formal education, in: 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), 2018, pp. 197–202, https://doi.org/10.1109/SISY.2018.8524609.
[17] S. Sohail, et al., Decoding ChatGPT: a taxonomy of existing research, current challenges, and possible future directions, Journal of King Saud University-Computer and Information Sciences (2023) 101675, https://doi.org/10.1016/j.jksuci.2023.101675.
[18] J. Hu, et al., Explainable AI models for predicting drop coalescence in microfluidics device, Chem. Eng. J. 481 (2024) 148465, https://doi.org/10.1016/j.cej.2023.148465.
[19] Y. Jiang, X. Yang, T. Zheng, Make chatbots more adaptive: dual pathways linking human-like cues and tailored response to trust in interactions with chatbots, Comput. Hum. Behav. 138 (2023) 107485, https://doi.org/10.1016/j.chb.2022.107485.

[20] P.P. Ray, ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154, https://doi.org/10.1016/j.iotcps.2023.04.003.
[21] K. Wach, et al., The dark side of generative artificial intelligence: a critical analysis of controversies and risks of ChatGPT, Entrepreneurial Business and Economics Review 11 (2) (2023) 7–30, https://doi.org/10.15678/EBER.2023.110201.
[22] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, F. Herrera, Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, https://doi.org/10.1016/j.inffus.2023.101805.
[23] A. Goossens, J. Vanthienen, Integrating GPT-technologies with decision models for explainability, in: World Conference on Explainable Artificial Intelligence, 2023, pp. 428–448, https://doi.org/10.1007/978-3-031-44070-0_22.
[24] M. Javaid, A. Haleem, R.P. Singh, ChatGPT for healthcare services: an emerging stage for an innovative perspective, BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (1) (2023) 100105, https://doi.org/10.1016/j.tbench.2023.100105.
[25] V.K. Singh, P. Singh, M. Karmakar, J. Leta, P. Mayr, The journal coverage of Web of Science, Scopus and Dimensions: a comparative analysis, Scientometrics 126 (2021) 5113–5142, https://doi.org/10.1007/s11192-021-03948-5.
[26] X. Hu, A. Liu, Y. Dai, Combining ChatGPT and knowledge graph for explainable machine learning-driven design: a case study, J. Eng. Des. (2024) 1–23, https://doi.org/10.1080/09544828.2024.2355758.
[27] K. Thapliyal, M. Thapliyal, Chatbot-XAI? The new age artificial intelligence communication tool for E-commerce, Role of Explainable Artificial Intelligence in E-Commerce 1094 (2024) 77–100, https://doi.org/10.1007/978-3-031-55615-9_6.
[28] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, S. Ananiadou, Towards interpretable mental health analysis with large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 6056–6077, https://doi.org/10.18653/v1/2023.emnlp-main.370.
[29] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, A. Holzinger, On generating trustworthy counterfactual explanations, Inf. Sci. 655 (2024) 119898, https://doi.org/10.1016/j.ins.2023.119898.
[30] P. Kieseberg, E. Weippl, A.M. Tjoa, F. Cabitza, A. Campagner, A. Holzinger, Controllable AI-an alternative to trustworthiness in complex AI systems?, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2023, pp. 1–12, https://doi.org/10.1007/978-3-031-40837-3_1.
[31] S. Rodriguez, J. Thangarajah, Explainable agents (XAg) by design, in: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, 2024, pp. 2712–2716. https://dl.acm.org/doi/10.5555/3635637.3663263.
[32] L. Wang, S. Zhang, Y. Wang, E.-P. Lim, Y. Wang, LLM4Vis: explainable visualization recommendation using ChatGPT, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2023, pp. 675–692, https://doi.org/10.18653/v1/2023.emnlp-industry.64.
[33] S. Lakhanpal, A. Gupta, R. Agrawal, Leveraging explainable AI to analyze researchers' aspect-based sentiment about ChatGPT, in: Intelligent Human Computer Interaction. IHCI 2023. Lecture Notes in Computer Science, 14532, 2024, pp. 77–100, https://doi.org/10.1007/978-3-031-53830-8_29.
[34] S. Krause, F. Stolzenburg, Commonsense reasoning and explainable artificial intelligence using large language models, in: European Conference on Artificial Intelligence, 2024, pp. 302–319, https://doi.org/10.1007/978-3-031-50396-2_17.
[35] S. Khambampati, S. Dondapati, T.V. Kattamuri, R.K. Pathinarupothi, CureNet: improving explainability of AI diagnosis using custom large language models, in: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), 2023, pp. 1–7, https://doi.org/10.1109/SMARTGENCON60755.2023.10442356.
[36] K. Kitamura, M. Irvan, R. Shigetomi Yamaguchi, XAI for medicine by ChatGPT code interpreter, in: Proceedings of the 2023 5th International Conference on Big-Data Service and Intelligent Computation, 2023, pp. 28–34, https://doi.org/10.1145/3633624.3633629.
[37] T. Susnjak, Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT, Int. J. Artif. Intell. Educ. 34 (2023) 452–482, https://doi.org/10.1007/s40593-023-00336-3.
[38] A. Ukil, J. Gama, A.J. Jara, L. Marin, Knowledge-driven analytics and systems impacting human quality of life- neurosymbolic AI, explainable AI and beyond, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 5296–5299, https://doi.org/10.1145/3583780.3615300.
[39] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, H. Müller, AI for life: trends in artificial intelligence for biotechnology, N. Biotech. 74 (2023) 16–24, https://doi.org/10.1016/j.nbt.2023.02.001.
[40] I. DeAndres-Tame, et al., How good is ChatGPT at face biometrics? A first look into recognition, soft biometrics, and explainability, IEEE Access 12 (2024) 34390–34401, https://doi.org/10.1109/ACCESS.2024.3370437.
[41] Y.H. Sung, S.Y. Chien, F. Yu, Preliminary validation of explainable AI interfaces across heuristics and information transparency, in: 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS), 2024, pp. 1–5. https://nccur.lib.nccu.edu.tw/handle/140.119/151148.
[42] V.S. Anoop, C.S. Krishna, U.H. Govindarajan, Graph embedding approaches for social media sentiment analysis with model explanation, International Journal of Information Management Data Insights 4 (1) (2024) 100221, https://doi.org/10.1016/j.jjimei.2024.100221.
[43] K. Demertzis, K. Rantos, L. Magafas, C. Skianis, L. Iliadis, A secure and privacy-preserving blockchain-based XAI-justice system, Information 14 (9) (2023) 477, https://doi.org/10.3390/info14090477.
[44] P. Berger, J.V. Garrel, How to design a value-based Chatbot for the manufacturing industry: an empirical study of an internal assistance for employees, Künstliche Intell 37 (2023) 203–211, https://doi.org/10.1007/s13218-023-00817-6.
[45] K.G. Barman, N. Wood, P. Pawlowski, Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use, Ethics Inf. Technol. 26 (3) (2024) 47, https://doi.org/10.1007/s10676-024-09778-2.
[46] R. Joshi, J. Graefe, M. Kraus, K. Bengler, Exploring the impact of explainability on trust and acceptance of conversational agents–A wizard of oz study, in: International Conference on Human-Computer Interaction, 2024, pp. 199–218, https://doi.org/10.1007/978-3-031-60606-9_12.
[47] Q. Yang, M. Ongpin, S. Nikolenko, A. Huang, A. Farseev, Against opacity: explainable AI and large language models for effective digital advertising, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 9299–9305, https://doi.org/10.1145/3581783.3612817.
[48] N.A. Etori, M. Gini, WisCompanion: integrating the socratic method with ChatGPT-based AI for enhanced explainability in emotional support for older adults, in: International Conference on Human-Computer Interaction, 2024, pp. 179–198, https://doi.org/10.1007/978-3-031-60606-9_11.
[49] M. Ramachandran, Quality framework for explainable artificial intelligence (XAI) and machine learning applications, in: Explainable Artificial Intelligence (XAI): Concepts, Enabling Tools, Technologies and Applications, 2023, pp. 115–138, https://doi.org/10.1049/PBPC062E_ch7.
[50] A. Wijekoon, et al., CBR driven interactive explainable AI, in: International Conference on Case-Based Reasoning Research and Development. ICCBR 2023. Lecture Notes in Computer Science, 14141, 2023, pp. 169–184, https://doi.org/10.1007/978-3-031-40177-0_11.
[51] X. Qu, H. Liu, Z. Sun, X. Yin, Y.S. Ong, L. Lu, Z. Ma, Towards building voice-based conversational recommender systems: datasets, potential solutions and prospects, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2701–2711, https://doi.org/10.1145/3539618.3591876.
[52] A. Banafa, Transformative AI: Responsible, Transparent, and Trustworthy AI Systems, CRC Press, 2024, https://doi.org/10.1201/9781032669182.
[53] T. Kokuti, L. Balazs, I. Andras, M. Rajcsanyi-Molnar, Collaborating with artificial intelligence - AI in business communication education, in: CANDO-EPE 2023 - Proceedings: IEEE 6th International Conference and Workshop Obuda on Electrical and Power Engineering, 2023, pp. 287–293, https://doi.org/10.1109/CANDO-EPE60507.2023.10417987.
[54] N.R. Haddaway, B. Macura, P. Whaley, A.S. Pullin, ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps, Environ. Evid. 7 (2018) 7, https://doi.org/10.1186/s13750-018-0121-7.