Computer Science Review 52 (2024) 100632

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions

Avyay Casheekar, Archit Lahiri, Kanishk Rath, Kaushik Sanjay Prabhakar, Kathiravan Srinivasan *

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

ARTICLE INFO

Keywords:
Computational intelligence
Artificial intelligence
Chatbots
Conversational agents
ChatGPT

ABSTRACT

This review paper offers an in-depth analysis of AI-powered virtual conversational agents, specifically focusing on OpenAI's ChatGPT. The main contributions of this paper are threefold: (i) an exhaustive review of prior literature on chatbots, (ii) a background of chatbots including existing chatbots/conversational agents like ChatGPT, and (iii) a UI/UX design analysis of prominent chatbots. Another contribution of this review is the comprehensive exploration of ChatGPT's applications across a multitude of sectors, including education, business, public health, and more. This review highlights the transformative potential of ChatGPT, despite the challenges it faces such as hallucination, biases in training data, jailbreaks, and anonymous data collection. The review paper then presents a comprehensive survey of prior literature reviews on chatbots, identifying gaps in the prior work and highlighting the need for further research in areas such as chatbot evaluation, user experience, and ethical considerations. The paper also provides a detailed analysis of the UI/UX design of prominent chatbots, including their conversational flow, visual design, and user engagement. The paper also identifies key future research directions, including mitigating language bias, enhancing ethical decision-making capabilities, improving user interaction and personalization, and developing robust governance frameworks. By solving these issues, we can ensure that AI chatbots like ChatGPT are used responsibly and effectively across a broad variety of applications. This review will be a valuable resource for researchers and practitioners in understanding the current state and future potential of AI chatbots like ChatGPT.

1. Introduction

Chatbots and conversational agents driven by Artificial Intelligence (AI) are software programmes or virtual assistants created to emulate human-like dialogues, engaging in interactive and lifelike conversations with users. These intelligent systems leverage AI capabilities to comprehend and react to user inputs, delivering an immersive and interactive conversational encounter [1–6].

A general interpretation of the working of a chatbot is displayed in Fig. 1 wherein, the user asks the chatbot "What is the weather in Bangalore?". The chatbot analyzes the user query and extracts the intent "to know the weather" and the entity "Bangalore". The chatbot then comprehends the user query by using its knowledge of the world to understand that the user is asking about the weather in Bangalore. Finally, the chatbot generates a coherent response in natural language, which is "The weather in Bangalore is currently 23 °C with chances of rainfall".

Chatbots can be implemented in various forms, including text-based chat interfaces, voice assistants [1–7], or even physical robots [7]. The underlying technology behind chatbots and AI conversational agents involves Natural Language Processing (NLP) and Deep Learning (DL). NLP enables the system to understand and interpret human language, allowing it to comprehend user queries, commands, or statements [1–3,5–7]. It involves tasks like text classification, entity recognition [3], sentiment analysis [6,7], and language generation [2,5,7,8]. We look at these applications in depth in Table 1.

The general architecture of a Chatbot in Fig. 2 shows the basic components and steps involved in the process of a chatbot understanding and responding to a user query. The illustration is structured into three primary parts: the module for Natural Language Understanding (NLU), the Dialogue Manager, and the module for Natural Language Generation (NLG) [9].

* Corresponding author.
E-mail addresses: avyaycasheekar.m2020@vitstudent.ac.in (A. Casheekar), archit.lahiri2021@vitstudent.ac.in (A. Lahiri), kanishk.rath2020@vitstudent.ac.in (K. Rath), kaushiksanjay.prabha2020@vitstudent.ac.in (K.S. Prabhakar), kathiravan.srinivasan@vit.ac.in (K. Srinivasan).

https://doi.org/10.1016/j.cosrev.2024.100632
Received 24 July 2023; Received in revised form 31 March 2024; Accepted 4 April 2024
Available online 9 April 2024

The NLU module is responsible for understanding the user's query. It does this by breaking down the query into its component parts, such as the words, phrases, and punctuation. It also identifies the intent of the query, such as whether the user is asking a question, making a request, or providing information. The Dialogue Manager is in charge of controlling the chatbot's dialogue with the user. It records the current status of the discussion, including what the user has asked and what the chatbot has answered in response. It also uses this information to decide how to respond to the user's next query. The NLG module assumes the responsibility of formulating a suitable response to the user's inquiry. It does this by using the chatbot's knowledge base and the Dialogue Manager's state information to create a response that is relevant to the user's query and that is understandable by the user.

Deep Learning also plays a crucial role in chatbots' ability to learn and improve over time. Initially, chatbots are trained on large datasets or predefined conversational flows to understand various intents and possible user inputs. They learn patterns, correlations, and context from the data to generate appropriate responses [1–8].

The origins of chatbots can be found in the middle of the 20th century, when computer scientists first started investigating the idea of interactive communication between humans and machines. The first primitive chatbot, known as ELIZA, was developed in the 1960s by Joseph Weizenbaum [10]. ELIZA used pattern matching techniques to simulate conversations by reflecting user input and asking clarifying questions [1–8].

In the following decades, chatbot development progressed with the introduction of rule-based systems [1]. These early chatbots followed predefined rules and decision trees to generate responses based on keyword matching. A prominent illustration of this is Artificial Linguistic Internet Computer Entity also known as ALICE developed by Richard Wallace during the mid-1990s. [11]. The emergence of increasingly sophisticated chatbots can be attributed to the significant advancements in the fields of NLP and Machine Learning (ML) during the 2000s. These remarkable progressions in NLP and ML techniques have paved the way for the development of highly refined chatbot systems. These AI-powered agents were able to comprehend meaning and produce responses resembling those generated by humans. The introduction of neural networks and deep learning algorithms significantly improved the quality and fluency of chatbot interactions [2–4,6,8].

In recent years, with the advent of cloud computing, big data, and the availability of large-scale datasets, chatbot development has accelerated. Companies like Facebook, Google, and OpenAI have made significant contributions to the field, introducing chatbot platforms and language models such as ChatGPT [12] which operate on the foundation of the Generative Pre-trained Transformer also known as GPT [13]. These models harness substantial volumes of data and sophisticated algorithms to provide more accurate and context-aware responses [1].

As chatbots continue to evolve, they are being integrated into various applications and platforms, including customer support, virtual assistants, and social media messaging [1,5,7,8]. Their applications in education, business, healthcare and other applications in computer science are discussed in depth in Section 4.

Looking ahead, the future of chatbots lies in the advancement of conversational AI, where chatbots will possess enhanced natural language understanding, emotional intelligence, and the ability to engage in more interactive and dynamic conversations [1,2,6,7]. With the growing adoption of chatbots across industries, the potential for these intelligent agents to revolutionize human–computer interactions and enhance user experiences is immense.

1.1. Shortcomings of prior literature

In our extensive analysis of past literature reviews on chatbots, we identified several deficiencies. These included a lack of analysis of chatbot User Interfaces and User Experiences (UI/UX), a lack of discussion of State-of-The-Art (SOTA) chatbot architectures, a lack of historical and evolutionary context for chatbots, a lack of discussion of chatbot applications in critical domains such as healthcare and finance, and a lack of discussion of the ethical and regulatory issues surrounding chatbots.

Our goal is to fill these gaps and provide researchers in this field with a comprehensive knowledge base on the current state of chatbot technology. Our review covers a detailed history and background of chatbots, popular datasets used in chatbot training, UI/UX analysis of prominent chatbots, and regulations and compliance involved with the data used to train chatbots.

Our analysis of research papers examine both the applications that the papers address and the impact and concerns of chatbot applications in depth. We take a domain-based approach, which shifts perspectives based on the respective domain environments. This allows us to consider all of the various perspectives on the important performance aspects of chatbots, as well as the areas that require the most attention across all domains. This domain-based view also provides a unique perspective on the state of chatbot applications and Large Language Models (LLMs) at present.

The open challenges identified and reviewed in this paper have not been discussed in other review papers. Previous review papers have been limited to the issues of chatbot accuracy and the limitations of survey methodology. While this approach is valid, it does not address the crucial aspect of the social readiness of chatbot technology, the harms it may cause, and how future research needs to deal with them, which is essential for wider use of this technology.

1.2. Contributions of this review

1. Extensive Review of prior literature on Chatbots
   This review is crucial because it lays the groundwork for our research and enables us to build upon existing knowledge. It helped identify the key themes, trends, and theories that have emerged, giving a comprehensive overview of the subject. Through our analysis of prior review work on chatbots, we have identified significant research gaps that warrant further investigation. These research gaps primarily pertain to the areas of UI/UX design analysis, regulations and compliance, as well as various avenues that require additional research within the field of chatbots. Apart from this, It provides a valuable resource for other researchers and practitioners interested in the topic, allowing them to access a comprehensive overview of the existing literature and state of chatbots(see Fig. 3).
2. Background of chatbots including existing chatbots/ conversational agents and ChatGPT
   In this review, we offer a detailed examination of the landscape of chatbots, with a particular emphasis on the role of virtual conversational agents and ChatGPT in particular. Our work draws crucial attention to the methods employed to enable normative conversation with ChatGPT agents, revealing their utility in various different contexts and increasing the understanding of their functionality.
3. UI/UX design analysis of prominent chatbots
   The examination of the UI/UX design across prominent chatbots is another significant contribution. By shedding light on the strengths and weaknesses of multiple designs, we offer crucial insights that help identify useful characteristics in design elements.
4. Survey of popular datasets used to train LLMs
   Our extensive survey of popular datasets used to train LLMs and other chatbots stands out as a crucial contribution to the review. In doing so, we emphasize the critical role of quality, diversity, and scale of data in shaping the capabilities of chatbots. We further explore issues with data solvency in other sections.

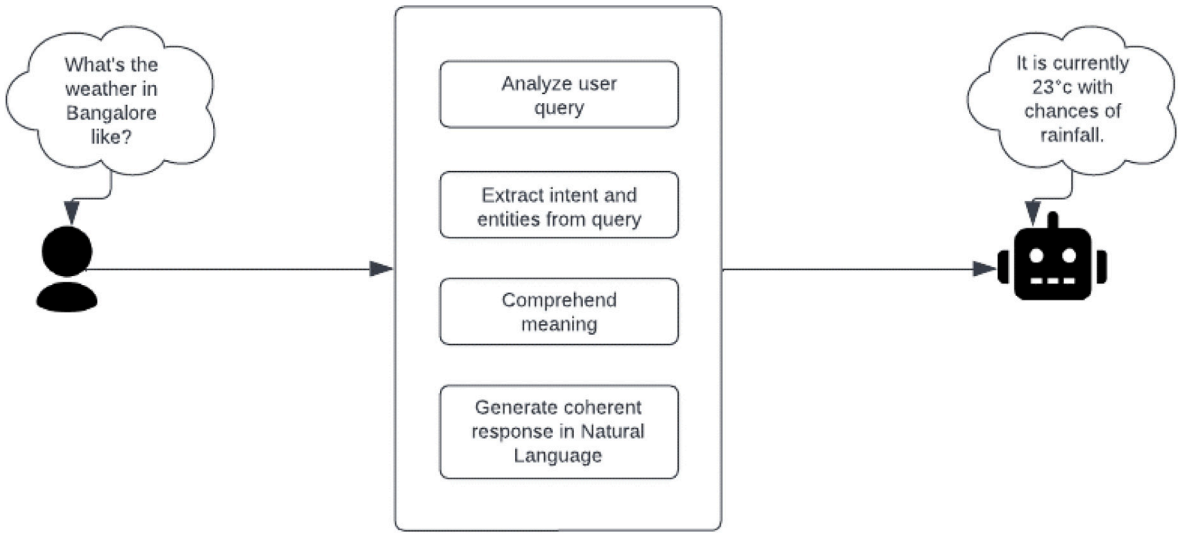

Fig. 1. AI-based chatbot – General interpretation.

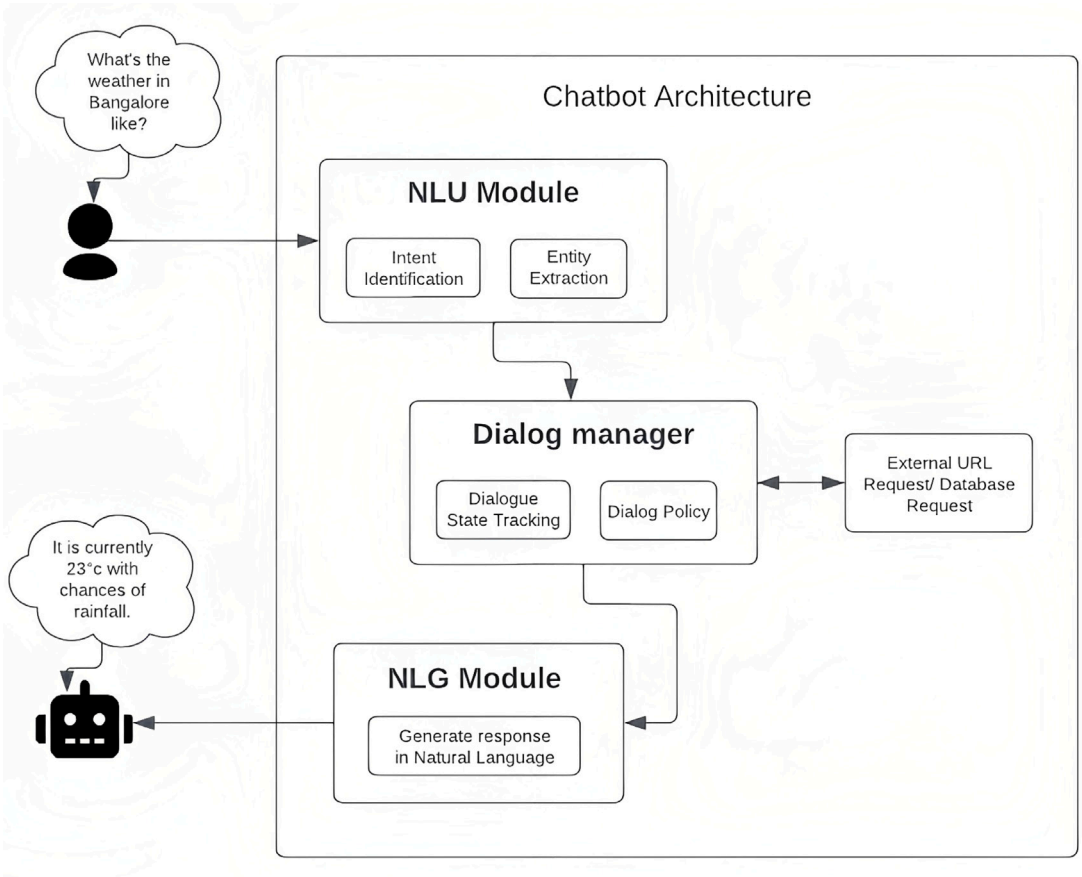

Fig. 2. AI-based chatbot – general architecture [9].

5. Discussion on the regulations and compliance involved with training data for chatbots and the ethical implications of AI
   Our discussion on regulations, compliance, and ethical implications of Chatbots, particularly with the training data for them, foregrounds a crucial but often overlooked aspect. As Chatbots such as ChatGPT continue to integrate into our everyday lives, understanding these issues is vital for protection against data malpractice and for the deployment and use of safe Chatbots.
6. In depth analysis of applications of Chatbots in important domains like education, business and health-care
   We conduct analysis and review on selected papers with a focus on applications being developed and their impact on various domains and such as Education, Business and Health-Care. Our findings were not only based on papers that made their own applications and models for NLP and ChatBot like functionalities, but also the deliberation and review of various papers that

Table 1
Comparison with previous reviews in Chatbot, AI-powered virtual conversational agents, ChatGPT.

Paper	Year	One-Phrase summary	Chatbots	Applications	Future directions	Shortcomings
Caldarini et al. [1]	2022	Reviews recent chatbot advances	Discusses chatbot evolution, methodologies, datasets, evaluation metrics, and implementation techniques.	Education and research; customer service in e-commerce.	Emphasizes need for universal chatbot evaluation framework.	Lacks HCI perspective and analysis in medical and financial domains.
Almansor and Hussain [8]	2020	Surveys chatbot approaches based on response generation.	Discusses chatbot history, classification, applications, and future direction.	Education and healthcare; gaming, retail, and media.	Need for standard quality measurement framework.	Lacks in-depth UI/UX analysis and detail on finance and e-commerce domains.
Park et al. [5]	2022	Reviews chatbot technology with focus on recent advancements.	Examines five chatbot technologies: data mining, semantic web, NLP, pattern matching, and context-aware computers.	Enterprise messaging, public services, entertainment services, and O2O and interactive commerce.	Future direction includes integrating voice command activation functionalities.	Lacks discussion of SOTA architectures and in-depth UI/UX analysis.
Singh and Beniwal [2]	2022	Surveys conversational agents covering approaches, frameworks, research questions, datasets, evaluation metrics.	Explores approaches, existing frameworks, datasets, evaluation metrics, and chatbot evolution.	Online customer service; hiring processes; patient queries; gadgets like Amazon Alexa, Siri, Google Home.	Need for more structured reviews, additional datasets and frameworks.	Lacks detailed exploration of chatbot applications and discussion of SOTA architectures.
Al et al. [3]	2021	Discusses chatbot evolution, types, current challenges, and future research.	Discusses chatbot history, roles in various fields, and emphasizes chatbots as complementary applications.	Education, banking, healthcare, tourism, and e-commerce.	Improve chatbot's ability to read user emotions and make conversations relevant.	Lacks in-depth UI/UX analysis and discussion of SOTA architectures.
Bilquise et al. [6]	2022	Analyses studies on the development of chatbots possessing emotional intelligence.	Examines the difficulties tackled, the strategies employed, and the assessment metrics applied in incorporating emotion into chatbot dialogues.	Surge in development of chatbots for customer service, education, healthcare, and social interaction.	Development of voice-based, domain-specific, and multimodal chatbots.	Does not discuss ethical implications of developing emotionally intelligent chatbots.
Adamopoulou and Moussiades [7]	2020	Presents the history, technology, and applications of chatbots.	Discusses chatbot history, weaknesses, implementation methodologies, use cases, and risks.	Education, customer service, healthcare, robotics, and food industry.	Improve language comprehension and production, integrate with other technologies, and research on ethical and social implications.	Lacks discussion of SOTA architectures and HCI analysis.
Hussain et al. [4]	2019	Investigates chatbots and conversational agents, as well as their categorization and design methodologies.	Discusses chatbot design techniques, conversation context handling, and current limitations.	Non-task-oriented conversations, customer service, virtual assistants, and language learning.	Enable chatbots to access online resources and make interaction style more human-like.	Lacks detailed history and applications of chatbots, and HCI analysis.
Bernardini et al. [14]	2018	Analyzes the state of the art of chatbot research.	Examines search trends for chatbots, identifies productive authors, institutions, and countries.	Potential to change industries but does not elaborate.	Intensive use of neural networks and research on ethical considerations.	Lacks sufficient detail on how chatbots could change industries.

expanded on the discussion about the implications of the use of ChatBots and LLMs in said fields.

7. Identify and review the literature on the most prevalent and relevant issues of Chatbots. Specifically, this paper analyzes the challenges of hallucinations, bias, and jailbreaks, which are major barriers to chatbot social acceptability and user-friendliness, and contribute to misinformation, and the harms associated with misalignment. The paper also reviews the existing literature on the causes of these challenges and the current state-of-the-art solutions that exist to counter them. Furthermore, the paper identifies and evaluates areas where even current state-of-the-art falls short.

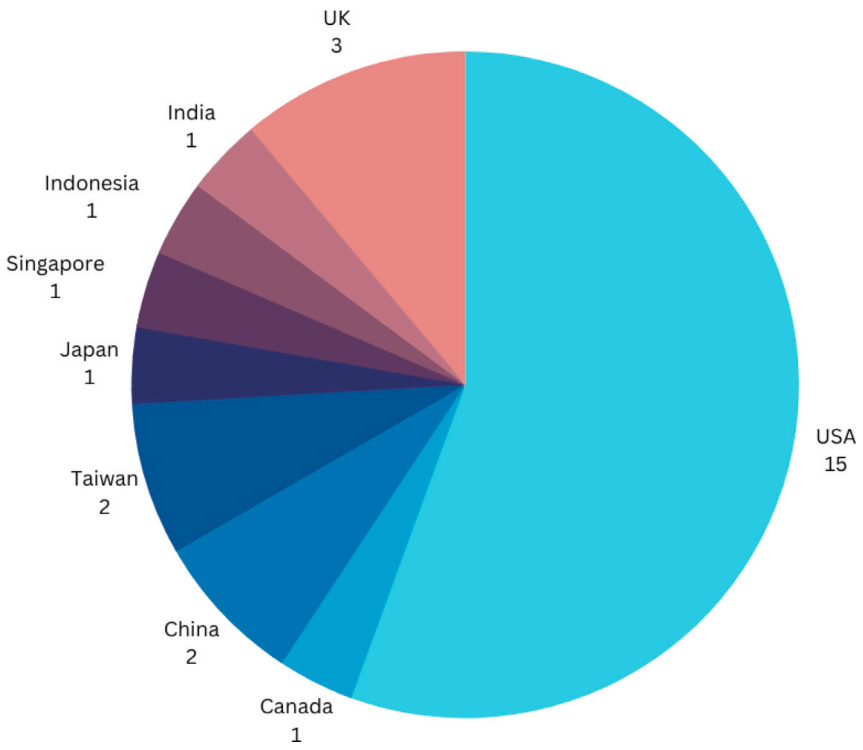

Fig. 3. Distribution of Review Papers on Chatbots across the world [15].

8. Recommendation for future research on chatbots.
   Our final contribution in this paper is presenting two perspectives on the future work that needs to be done on ChatBots and LLMs.

1.3. Organization of this review

This review has been structured into 7 main sections namely Introduction, Survey Methodology, Background of Chatbots, Applications of Chatbots, Open Challenges, Future Research Directions and Conclusions. The organization of this review is detailed in Fig. 4.

2. Materials and methods

This section goes over the planning, strategies and process of filtration of research papers for selection.

2.1. Planning the review

The paper review and selection process was guided by the given list of steps to be followed to first collect and filter out papers to be considered for review.

1. Search strategy: Strategise as to what papers need to be included, what to look for in terms of keyword selection and filtering, and where to look for them;
2. Inclusion criteria: Define the criteria to approve of or deny papers for reviewing;
3. Quality Evaluation: Define the way in which a paper is to be evaluated for its research value and relevancy to the paper's research direction studies, and filter them based on selection criteria;

2.2. Search strategy

For this review, advances on Chatbot, AI-powered virtual conversational agents, ChatGPT and any other work on LLMs or NLP were queried in databases including Google Search, Google Scholar, Arxiv and Science Direct. The prevailing trend within the field of LLM has been the publication and dissemination of details regarding LLMs, such as LlaMa and Meta, primarily through articles that are submitted to archives. Consequently, we have incorporated papers sourced from archives into our research. The papers should ideally be published between January of 2010, to March of 2024.

Given below is the search string used

("Chatbot" OR "AI-powered virtual conversational agents" OR "GPT" OR "ChatGPT" OR "BeRT" OR "LLMs") AND ("Application" OR "Use Case") AND ("Specific Domain Name : Education, Healthcare, Medical, Business, Miscellaneous") (see Fig. 5)

For the given papers, additional papers were considered that provided the base context of concepts the reviewed papers were based on.

2.3. Selection criteria

The listed points describe the details taken into account for the first step of filtration after gathering papers from the query string described before,

1. Inclusion Criteria: Research papers published between January 2017 till March 2024, the date of writing, on Chatbots, AI-powered virtual conversational agents and related applications were sought after. The papers should be English, and make use of commercially available LLMs such as ChatGPT, GPT 3, GPT 4, BeRT or their own LLM, or have conducted a review on the same. Recent researches and works were favoured more. Additional papers that conducted similar surveys and provided base context for the domains being researched were also included. Papers

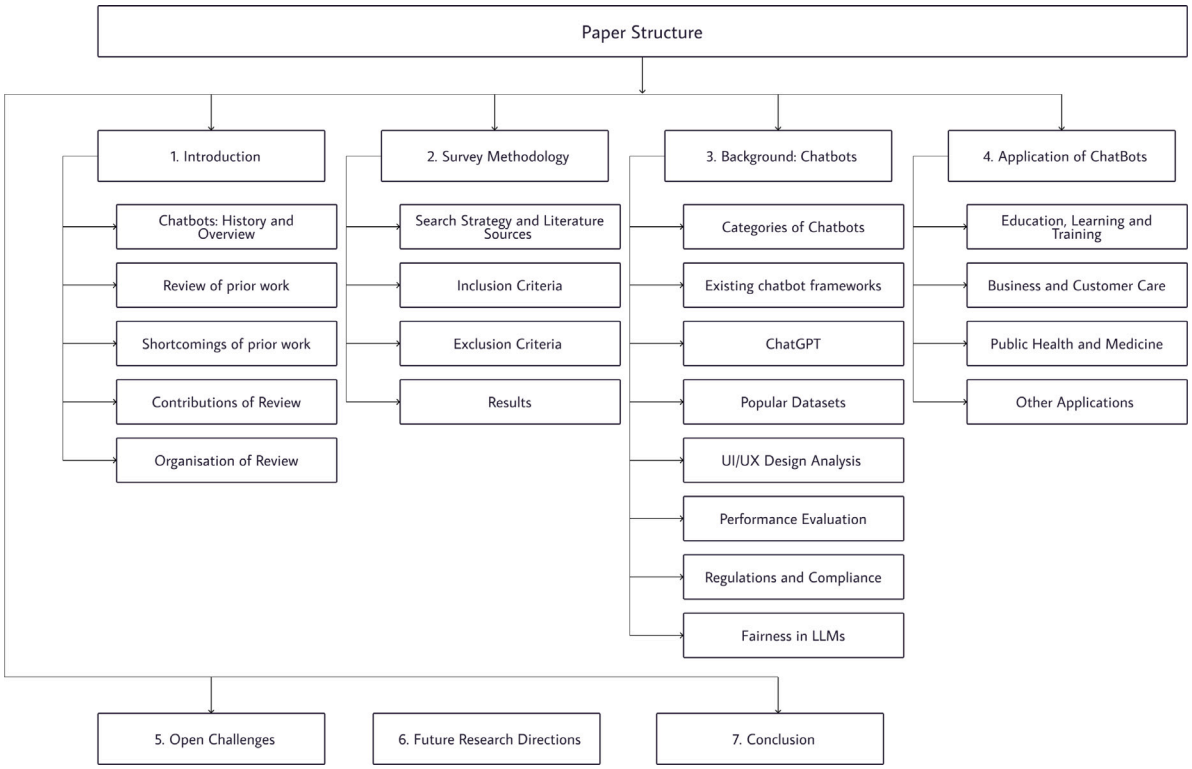

Fig. 4. Organization of this review.

SEARCH STRING

("Chatbot" OR "AI-powered virtual conversational agents" OR "GPT" OR "ChatGPT" OR "BeRT" OR "LLMs") AND ("Application" OR "Use Case") AND ("{Specific Domain Name : Education, Healthcare, Medical, Business, Miscellaneous}")

Fig. 5. Search String used to pool together relevant papers.

written before January 2017 if they pertain to well known concepts and technology commonly used in the field.

2. Exclusion Criteria: Papers not written in English, reported before January 2017 were excluded. Opinions, abstract only reports, papers without abstract, dissertation and thesis, practice guidelines, commentaries, editorials, books, letters, works not related to ChatBots, LLMs, NLP and surrounding technology were all excluded from this review.

2.4. Quality assessment

Our Quality Assessment procedure were carried out as follows,

1. Duplicate removal: Removing duplicate works that were found in the list of selected papers
2. Impurity removal: Removing works that were no relevant to the topic being reviewed
3. Filter by content of abstract: Analysing the content of the abstract of the papers to determine relevancy of paper
4. Filter complete content of papers: Read through full articles for final review.

With the papers being filtered out to only ones that were relevant to our review, we carried out further qualitative assessments, such as assessing papers on

1. If the paper had a novel proposed methodology, the architecture used was properly explained, and the reasoning behind it was justified. Metrics were displayed highlighting the performance of the proposed architecture.
2. If the paper was a review paper, the direction of the survey and research objectives were properly described, and the conclusion was related to these research objectives.

2.5. Results

Fig. 6 illustrates the process of our quality assessment for this paper, while tallying the number of papers considered at each stage. Fig. 7 illustrates the year of publishing of the papers finalised for this review.

3. Background - chatbots, AI-powered virtual conversational agents, ChatGPT

A chatbot is a software application engineered to mimic human conversation. These chatbots can leverage AI and NLP technologies to comprehend and produce responses that closely resemble human interaction [16]. AI-powered virtual conversational agents are chatbots that use AI to provide more advanced and personalized responses. ChatGPT is an example of an AI-powered virtual conversational agent. It was developed by OpenAI and uses deep learning algorithms and NLP to understand and generate responses to user input [17]. ChatGPT marks a notable progression in the field of conversational AI, crafting responses that emulate human dialogue when given natural language inputs. While ChatGPT has the capability to formulate responses that might be leveraged by a chatbot or virtual assistant [18], it does not possess the full suite of features necessary to act as a standalone conversational AI platform [19]. This means that while ChatGPT can provide advanced and personalized responses, it still requires additional features and capabilities provided by a conversational AI platform in order to function effectively.

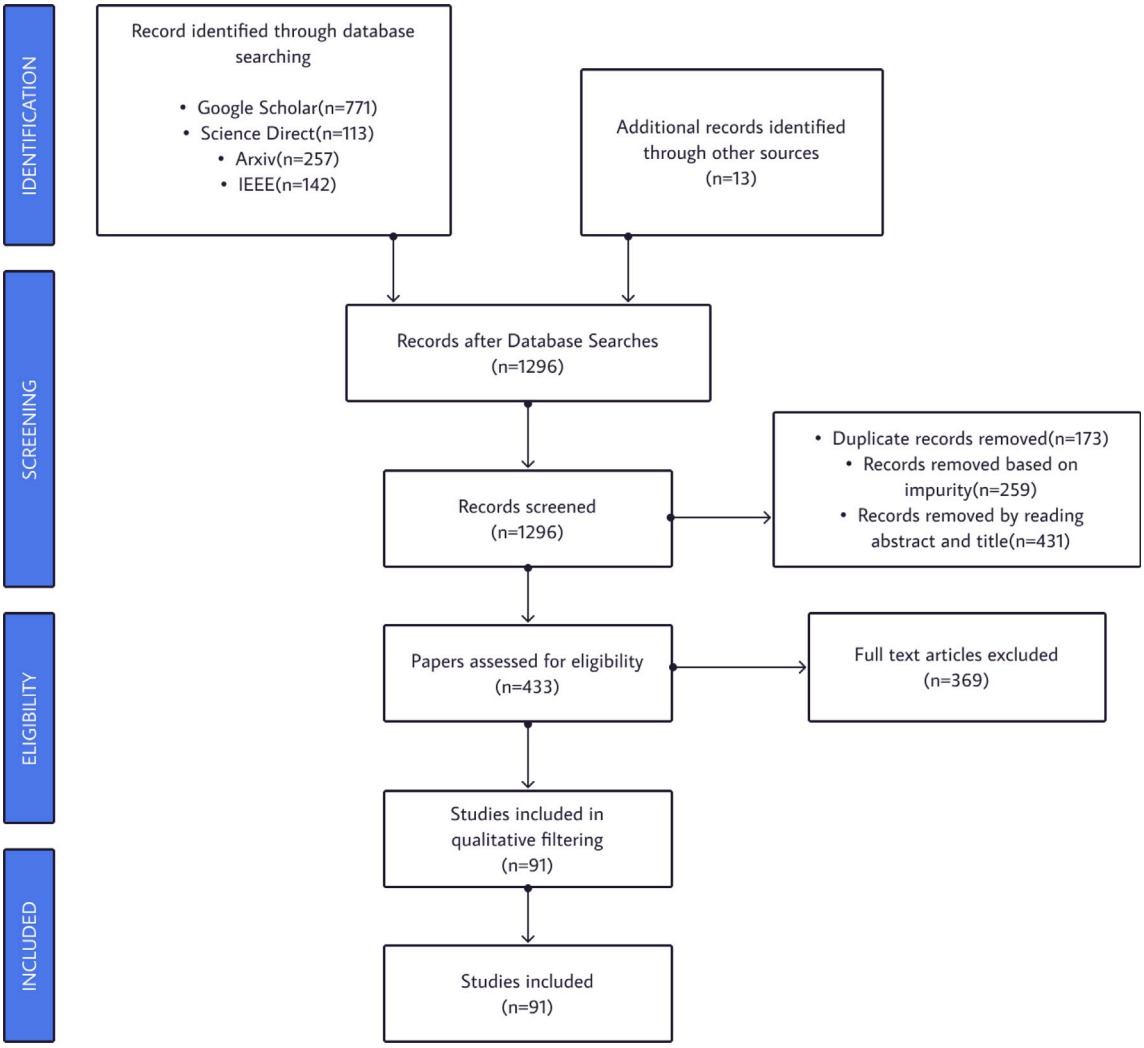

Fig. 6. Filtration process and paper count visualized throughout the selection process using a flowchart.

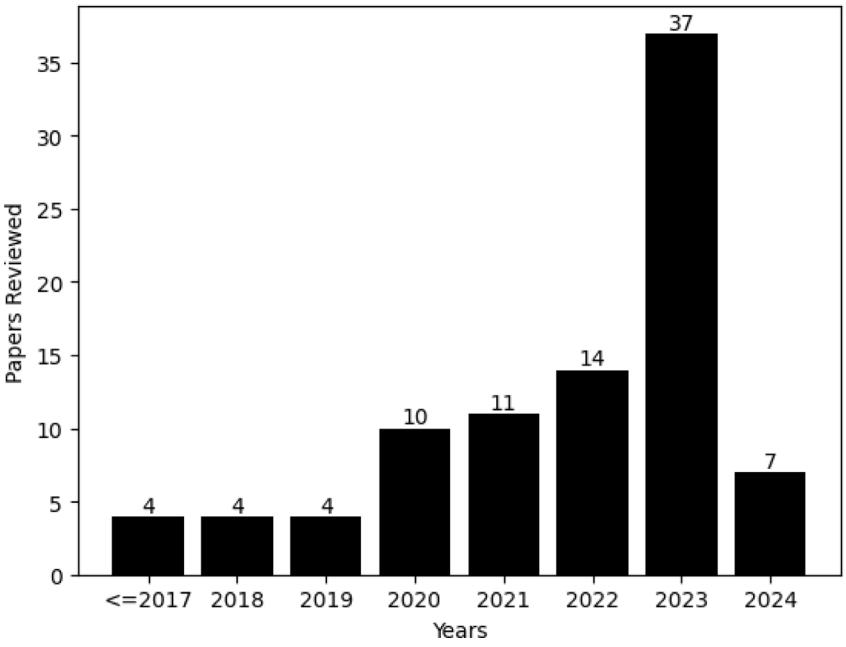

Fig. 7. Histogram on publications by year.

We aim to look at some methods employed to enable normative conversation with ChatGPT agents. We will categorize them and look at the use cases in detail.

3.1. Categories of chatbots and AI-powered virtual conversational agents

We classify chatbots (see Fig. 8) into the following categories:

1. **Task-oriented chatbots:** These chatbots are engineered to execute specific tasks for users, such as flight booking, food ordering, or weather checking. They typically operate within a predefined dialogue flow and a confined domain of knowledge. Depending on their approach to user input and response generation, they can be rule-based or data-driven.
    (a) **Personal assistant chatbots:** Personal assistant chatbots aim to aid users in managing their personal tasks, such as appointment scheduling, reminder setting, or product ordering. They typically communicate in a friendly and conversational tone and can integrate with various platforms and services.
    (b) **Customer service chatbots:** Customer service chatbots are created to offer users support and assistance for a specific product or service. They typically communicate in a professional and informative tone and can handle common queries, complaints, or feedback.
2. **Social chatbots:** Social chatbots are created to engage users in casual conversations for entertainment or emotional support. They strive to emulate human-like interactions and personalities. They typically rely on data-driven methods, such as neural networks or generative models, to learn from large corpora of conversational data.
    (a) **Entertainment chatbots:** Entertainment chatbots aim to offer users fun and engaging experiences, such as games, quizzes, jokes, or stories. They typically communicate in a creative and humorous tone and can interact with users' emotions or preferences.
    (b) **Social good chatbots:** Social good chatbots aim to offer users social benefits or awareness, such as health advice, education resources, or charity donations. They typically communicate in a positive and motivational tone and can inspire users to take action or learn something new.
3. **Knowledge-based chatbots:** Knowledge-based chatbots aim to offer users factual information from various sources, such as Wikipedia, news articles, or databases. They employ techniques of natural language understanding and natural language generation to interpret user inquiries and generate pertinent responses.
    (a) **Informational chatbots:** Informational chatbots aim to offer users relevant information from various sources, such as news, weather, sports, or entertainment. They typically communicate in a neutral and factual tone and can answer questions or deliver updates.
    (b) **Educational chatbots:** Educational chatbots aim to offer users learning opportunities from various domains, such as languages, mathematics, science, or history. They typically communicate in an interactive and adaptive tone and can teach concepts or test skills.
4. **Hybrid chatbots:** Hybrid chatbots are engineered to combine the features of task-oriented, social, and knowledge-based chatbots. They can perform multiple functions for users depending on their needs and preferences. They use a combination of rule-based and data-driven methods to handle complex dialogues and scenarios.
    (a) **Conversational AI agents:** Conversational AI agents aim to offer users personalized and proactive services across multiple channels, such as voice, text, email, social media etc. They typically communicate in an intelligent and natural tone and can understand context, sentiment, intent etc.
    (b) **Virtual assistants:** Virtual assistants aim to offer users comprehensive assistance for various aspects of their lives, such as work, lifestyle, family etc. They typically communicate in an empathetic and supportive tone and can integrate with multiple devices, services, data sources etc.

3.2. Existing AI-powered virtual conversational chatbot frameworks

Our review indicates a distinct progression towards advanced and intricate language generation abilities. Since the groundbreaking work of IBM Watson Assistant in 2017, there has been a considerable evolution towards transformer-based models by 2023, including the likes of GPT-2, GPT-3, GPT-4, ChatGPT, CLAUDE, and BARD. These models have demonstrated an ability to generate coherent and contextually relevant sentences, thereby enhancing the quality of human–computer interaction. Furthermore, the training methodologies have evolved, with recent models like GPT-4 and ChatGPT employing Reinforcement Learning from Human Feedback (RLHF) to reduce harmful and untruthful outputs. The services provided by these chatbots have also diversified, ranging from writing assistance to creative writing, and from drafting emails to answering questions [20–22]. We explore several frameworks of chatbots in detail in Table 2.

In Fig. 9, we present a timeline showcasing the release years of various AI-powered virtual conversational chatbot platforms, ranging from IBM Watson Assistant's introduction in 2017 to the debut of BARD, CLAUDE, and GPT-4 in 2023. This timeline provides a clear visual representation of the rapid progression in the field of AI chatbot platforms.

The importance of this table lies in its ability to provide a comprehensive overview of the progression of AI-powered conversational chatbot frameworks over the years. For a review paper, this table serves as a valuable resource to understand the evolution and current state of chatbot technology. The selected columns provide insights into the functionality, dialogue management, human-aid in training, and services provided by each chatbot platform, thereby offering a holistic view of the field. This information is crucial for identifying trends, understanding the strengths and weaknesses of different models, and predicting future developments.

3.3. ChatGPT

Introduced in November 2022, ChatGPT [12] is an AI chatbot developed by OpenAI. Its development involved the use of supervised and reinforcement learning techniques, and it is rooted in the GPT-3 family of large language models from OpenAI [21].ChatGPT can handle general and technical questions, engage in casual conversations, and manage multiple topics, contexts, intents, sentiments, and references within a dialogue.

ChatGPT generates text based on user queries using a neural network architecture called Generative Pre-trained Transformer (GPT). This architecture, consisting of multiple layers of self-attention mechanisms, learns from vast amounts of text data. GPT-3.5, an upgraded version of GPT-3 with additional parameters and better performance, was fine-tuned on chat data to get ChatGPT [26]. Both semi supervised learning and reinforcement learning techniques were used to train it [21,27] and supervised learning was used during Reinforcement Learning through Human Feedback(RLHF) [28]. In the process of supervised learning, the model was given dialogues in which trainers took on

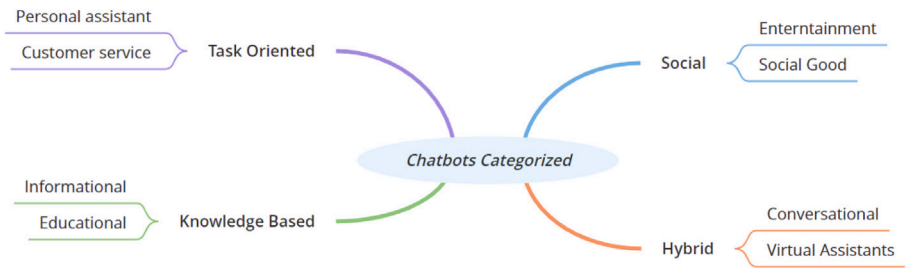

Fig. 8. Classification of different chatbot types.

Table 2
Existing AI-powered virtual conversational chatbot frameworks.

Chatbot platform	Year of release	Category	Functionality	Dialogue management	Human-aid	Service provided
BARD [23]	2023	AI-powered conversational chatbot	BARD is a chatbot powered by generative artificial intelligence, a product of Google's development efforts	Initially, BARD was built on the foundation of the LaMDA family of LLM, and subsequently, it was based on the PaLM LLM.	Contractors were used to train the initial model through human feedback	BARD is an experimental service provided by Google. It is available in English, Japanese, and Korean.
CLAUDE [24]	2023	AI-powered conversational chatbot	CLAUDE is an AI assistant, stemming from Anthropic's investigations into the development of AI systems that are beneficial, truthful, and devoid of harm.	The specific details about the dialogue management system used by CLAUDE are not readily available	The specific training approach for CLAUDE is not directly mentioned, but it is plausible to conjecture that it incorporates a blend of supervised learning and reinforcement learning methods.	CLAUDE was launched via an API and is available to people in the US and UK in open beta. It allows conversational abilities for users to interact with
GPT-4 [20]	2023	AI-powered conversational chatbot	GPT-4 has been trained on a wide array of internet text, but with the added capability of accessing external knowledge from a variety of sources	GPT-4 uses a transformer-based model to manage dialogue, considering the entire context of the conversation for generating responses	GPT-4 was trained using Reinforcement Learning from Human Feedback (RLHF), a method that helps to reduce harmful and untruthful outputs	GPT-4 is a powerful tool released by OpenAI, and has found applications in data analysis, language handling, code generation among other things
ChatGPT [12]	2022	AI-powered conversational chatbot	ChatGPT stands out for its ability to allow users to guide a conversation towards a preferred length, format, style, level of detail, and language	ChatGPT is built on the foundation of GPT-3.5 and GPT-4, which are part of OpenAI's proprietary series of GPT models	The enhancement of the model's performance was achieved through fine-tuning with the help of human trainers	Initially, ChatGPT was launched as a research preview available at no cost, but due to its widespread popularity, OpenAI has transitioned the service to a freemium model. It is optimized for chat-like inputs and has gained popularity for this feature
GPT-3 [21]	2020	AI-powered conversational chatbot	GPT-3 is a natural language processing AI model. Based on the information provided, it can produce text that resembles that of a human.	GPT-3 uses a transformer-based model to manage dialogue, considering the entire context of the conversation for generating responses	GPT-3 was trained on a diverse range of internet text. There was comparatively lesser human fine tuning done	GPT-3 has been utilized in many different applications, including creating emails and producing Python programmes.
GPT-2 [22]	2019	AI-powered conversational chatbot	GPT-2 is a large transformer-based language model capable of generating coherent and contextually relevant sentences	GPT-2 uses a transformer-based model to manage dialogue, considering the entire context of the conversation for generating responses	A variety of online material was used to train GPT-2. It does not, however, have access to any personal information about persons until specifically mentioned in the chat or know which exact papers were in its training set.	GPT-2 has been used in a variety of applications, from writing assistance to creative writing and more.
IBM Watson Assistant [25]	2017	Conversational AI platform	Exceptional customer care based on the advantage of AI	Machine learning, natural language understanding, and an integrated dialogue editor	Connects to the platform for contact centre customer data and other online assistance tools	Conversational interfaces into any application, device, or channel

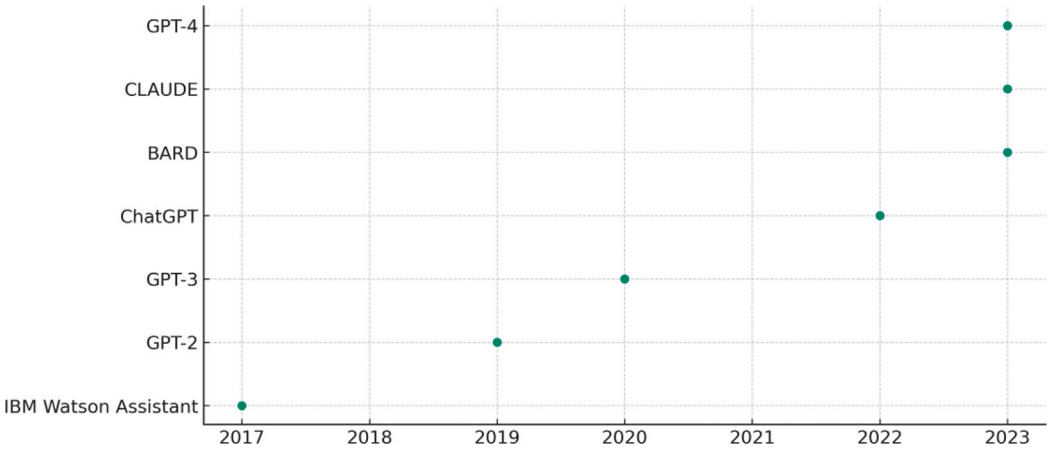

Fig. 9. Timeline of chatbots reviewed in this section.

the roles of both the user and the AI assistant. This allowed the model to imitate the tone and substance of the talks. In reinforcement learning, trainers ranked responses generated by different models according to their quality, providing feedback signals such as likes or dislikes. The model then optimized its reward function based on these signals using policy gradient methods.

ChatGPT(GPT3.5) was trained on a blend of text and code data from various web sources like Wikipedia, Reddit, Stack Overflow, GitHub, etc [21]. The data underwent preprocessing steps like tokenization, deduplication, filtering, and sharding to reduce noise and improve quality.

Performance evaluation of ChatGPT involved both automatic metrics and human feedback methods [29]. Automatic metrics, computed without human intervention, included measures like perplexity, diversity, coherence, and relevance. Human feedback methods involved subjective evaluations like rating, ranking, preference elicitation, and dialogue analysis [29].

Despite its impressive results, ChatGPT faces limitations and challenges such as factual errors, nonsensical answers, repetition, inconsistency, and bias [30]. Factual errors might arise from insufficient or outdated data sources, lack of external knowledge sources, or incorrect inference by ChatGPT. Nonsensical answers could be due to poor data quality, insufficient context awareness, or random generation by ChatGPT. Repetition, where ChatGPT produces similar or identical responses for different queries or within a conversation, might occur due to limited vocabulary size, low diversity score, or lack of memory by ChatGPT.

3.4. Datasets - AI-powered virtual conversational chatbot

In recent years, the development and utilization of AI-powered virtual conversational agents or chatbots have seen significant advancements, driven by the increasing availability of diverse and large-scale datasets. Table 3 provides an overview of some of the most influential datasets used in the training of these models. The WebText dataset, although not publicly available, played a crucial role in the training of OpenAI's GPT-2 model [22]. The Common Crawl dataset, a broad and diverse collection of web pages crawled from the internet, was instrumental in the training of the GPT-3 model [21]. The Pile dataset, a proprietary dataset comprising a diverse collection of texts from various sources, was also used in the training of the GPT-3 model. Lastly, the LLaMA models dataset, used to train Meta's LLaMA models, is a combination of the Common Crawl, Wikipedia, and C4 datasets [31]. These datasets, each with their unique characteristics and strengths, have contributed significantly to the advancements in the field of AI-powered virtual conversational agents.

Unlike most chatbot reviews that primarily focus on the technical capabilities, this section provides an in-depth examination of the datasets that underpin their abilities. This is crucial, as the quality, diversity, and scale of these datasets are key determinants of the models' performance and capabilities.

Our overview reveals a wide range of datasets utilized in the training of conversational agents. For instance, the WebText dataset, used in training OpenAI's GPT-2 model, and the Common Crawl dataset, used for GPT-3, highlight the importance of large-scale, diverse web text data in training these models. On the other hand, the proprietary Pile and LLaMA datasets demonstrate the potential of specialized, curated datasets in enhancing the performance of these models.

In Fig. 10, we present a combined bar and line chart showcasing the scale of datasets used for training various large language models. Each dataset, denoted along the x-axis, is represented by a pair of bars indicating the number of phrases and tokens respectively. The overlying line plot, with datapoints corresponding to each dataset, emphasizes the number of tokens in each dataset. Remarkably, the LLaMA models dataset stands out for its high token count.

This datasets overview section, therefore, not only provides researchers and practitioners with valuable insights into the data foundations of different conversational agents but also offers a clear direction for future dataset creation and curation.

3.5. UI UX design

In the realm of UI/UX design for AI-powered virtual conversational agents, a variety of approaches have been adopted by different platforms. OpenAI's GPT-2 and GPT-3, for instance, have been lauded for their high usability and integration capabilities, with GPT-3 in particular receiving praise for its advanced conversational abilities and creative text generation [21]. ChatGPT, also by OpenAI, was designed specifically for conversation and has been highly commended for its ability to generate meaningful, relevant responses. Microsoft's Cortana, while offering high usability through voice and text interactions and integration with Windows OS, has received mixed reviews, with some users finding it less intuitive than its competitors [37]. Google Assistant stands out for its very high usability and its ability to understand complex queries and provide useful, context-aware responses [38]. Facebook's Messenger based chatbots, while appreciated for their convenience, have been noted for their limitations in understanding complex queries [39]. Apple's Siri, despite being praised for its personality and voice recognition, has received mixed reviews, with some users finding it lacking in understanding context [40]. Amazon's Alexa, primarily a voice interaction platform, has been generally positively received, especially for its smart home control capabilities [38]. The diversity in these approaches underscores the ongoing evolution

Table 3
Datasets Used to Train Large Language Models.

Paper	Dataset	Open source/closed	Content type and source	Number of phrases	Number of tokens	Description
Radford et al. [22]	WebText	Open Source	Web pages scraped from outbound links on Reddit.	6,500,000	33 Billion	Used to train GPT-2 model.
Luccioni and Viviano [32]	Common Crawl	Open Source	Web pages crawled from the internet.	42 Billion	400 Billion	Used to train GPT-3 model.
Gao et al. [33]	Pile dataset (Proprietary)	Closed Source (Access granted by application)	A diverse collection of texts from various sources including books, scientific papers, and web pages.	800GB+ text data	400 Billion	Used to train GPT-3 model.
Touvron et al. [31]	LLaMA models dataset (Proprietary)	Closed Source (Access granted by application)	Common Crawl, Wikipedia, and C4 datasets were used to train LLaMA models.	1.5 billion phrases	1.4 Trillion	The Common Crawl, Wikipedia, and C4 datasets were used in combination by Meta to train its LLaMA models making this dataset
Rajpurkar et al. [34]	SQuAD	Open Source	Questions and answers derived from Wikipedia articles.	100000+ question-answer pairs	NA	Diverse collection of 100000+ questions and answers produced by humans through crowdsourcing.
Talmor et al. [35].	CommonSense QA	Open Source	Questions created by Amazon Mechanical Turk workers using ConceptNet	12,247 questions	NA	A dataset for common sense question answering generated using ConceptNet relations.
Bajaj et al. [36].	MS MARCO	Custom (research-only, non-commercial)	Question answering dataset featuring real Bing questions, natural language generation, passage ranking	Over 1 million queries	NA	A large-scale dataset focused on machine reading comprehension, question answering, and passage ranking.

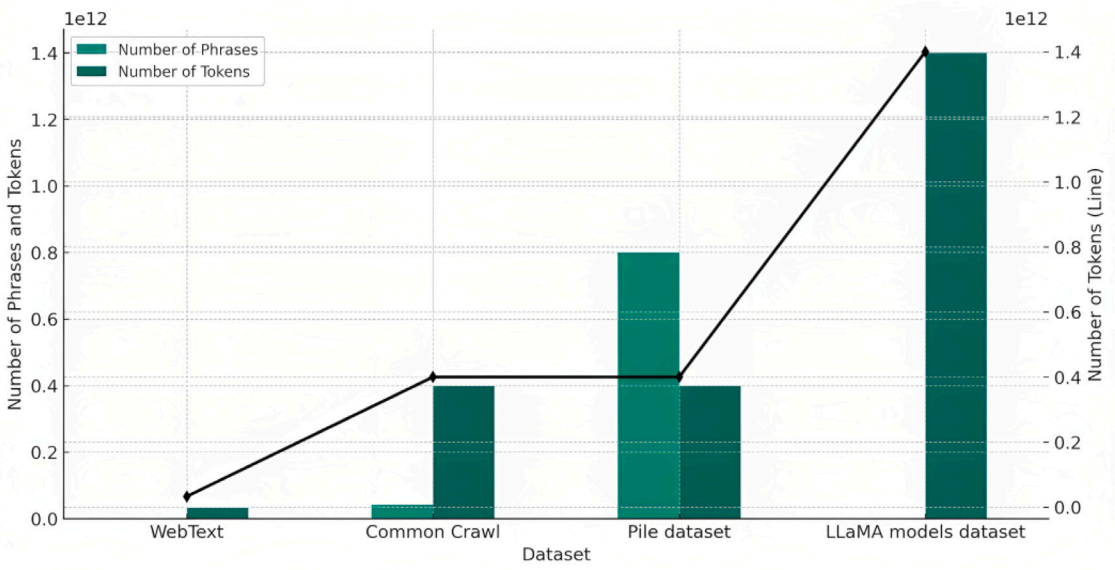

Fig. 10. Comparison of datasets used to train large language models.

and innovation in the field of AI-powered virtual conversational agents. We showcase a comparison of feedback in Table 4

The incorporation of a UI/UX review section in this paper is a novel and significant contribution to the literature on AI-powered conversational agents. Unlike traditional reviews, which often focus solely on technical capabilities, this section provides a comprehensive evaluation of the UI and UX offered by these agents. This is of paramount importance, as the UI and UX are critical determinants of user engagement and satisfaction.

Our review reveals a diversity of approaches in the design of conversational agents. For instance, Google Assistant and ChatGPT have been highly commended for their ability to understand complex queries and generate context-aware responses, demonstrating the importance of sophisticated natural language processing capabilities in enhancing UX. On the other hand, the convenience offered by Facebook Messenger's chatbots, despite their limitations in understanding complex queries, underscores the value of seamless integration into existing platforms in enhancing UI.

Table 4
User Feedback comparison across UI/UX Design, Usability, and Universal usability.

Chatbot	Authors and year	Usability	Universal usability	User feedback - UI/UX Design
GPT4	OpenAI, 2023 [20]	Very High: Designed specifically for conversation;	Very High: Easily integratable via API	Highly positive integration with plugins, codeinterpreter and other tools via API
Claude	Anthropic, 2023 [24]	Very High: Designed specifically for conversation;	Very High: Easily integratable with platforms such as Poe	Positive: Praised for long context handling
BARD	Google, 2023 [23]	High: Designed for large-scale model training; a single GPU is capable of training models with more than 13 billion parameters	Moderate: Requires access to GPU and knowledge of PyTorch for optimal use	Positive: praised for democratizing large-scale model training, enabling data scientists with access to just one GPU to use it
ChatGPT	OpenAI, 2022 [12]	Very High: Designed specifically for conversation; text interactions	High: Available for integration in various platforms	Highly positive, praised for its conversational abilities and ability to generate meaningful, relevant responses
GPT-3	OpenAI, 2020 [21,41]	High: Text interactions; more advanced and capable than GPT-2	High: Available for integration in various platforms	Highly positive, praised for its conversational abilities and creative text generation
GPT-2	OpenAI, 2018 [22,42]	High: Text interactions; used as a basis for many applications	High: Available for integration in various platforms	Generally positive, especially in terms of text generation capabilities, but not designed specifically for conversation
Cortana	Microsoft, 2016 [38]	High: Voice and text interactions; integrates with Windows OS	Moderate: Available on multiple platforms but primarily tied to Microsoft products	Generally positive, but some users find it less intuitive than competitors
Google Assistant	Google, 2016 [38]	Very High: Voice and text interactions; integrates with Google's suite of products	High: Available on multiple platforms including Android, iOS, Google Home	Highly positive, praised for its ability to understand complex queries and provide useful, context-aware responses
Siri	Apple, 2010 [38]	High: Voice and text interactions; integrates with iOS	Low: Primarily tied to Apple products	Mixed to positive: Praised for its personality and voice recognition, but some users find it lacking in understanding context
Alexa	Amazon, 2014 [38]	High: Primarily voice interactions; integrates with Amazon's suite of products	Moderate: Primarily tied to Amazon Echo devices, but has some cross-platform compatibility	Generally positive, especially praised for its smart home control capabilities

This UI/UX review section, therefore, not only provides users with valuable insights into the usability of different conversational agents but also offers developers a clear direction for future improvements. The best practices identified in this review, such as designing for complex query understanding, ensuring high usability and integration capabilities, and tailoring the design to specific use cases, can guide the development of future AI-powered conversational agents. This, in turn, can lead to the creation of more user-friendly and effective conversational agents, thereby advancing the field of AI-powered virtual conversational agents.

In Fig. 11, a comprehensive comparison of various chatbots, including GPT4, Claude, BARD, ChatGPT, GPT-3, GPT-2, Cortana, Google Assistant, Siri, and Alexa, is visualized using a bar plot. Each chatbot is evaluated across three key dimensions: Usability, Universal Usability, and User Feedback in UI/UX Design. It is clear that GPT4, Claude, and ChatGPT exhibit superior performance across all dimensions, highlighting their effectiveness in terms of usability and user satisfaction.

The rapid evolution of Large Language Models (LLMs) has ushered in a new era of AI personalization, with each model offering unique capabilities and fine-tuning features to cater to diverse user needs. Google's Gemini stands out for its exceptional multimodal understanding and long-context processing, enabling deep, nuanced interactions across various modalities. This level of sophistication allows for highly personalized user experiences, as the model can comprehend and respond to complex queries and instructions in a context-aware manner [43]. [44] looks at Personas for AI usage to increase personalization and make human like interaction of previously black box systems more understandable. We explore this in Table 5.

Anthropic's Claude 3 prioritizes safety and responsible AI, ensuring that personalization occurs within ethical and societal bounds [24,46]. By a robust ethical framework into its design, Claude 3 can generate content that aligns with individual preferences while adhering to established norms and values [43]. This approach is particularly crucial in domains such as healthcare, finance, and education, where trust and accountability are paramount.

OpenAI's GPT-4 excels in adaptability and performance, thanks to its fine-tuning mechanisms and ability to learn from human feedback [51]. This makes GPT-4 a versatile tool for personalization across a wide range of applications, from content creation to task automation [52]. However, as with all LLMs, it is essential to consider the potential limitations and biases that may arise from the training data and fine-tuning processes.

The LLaMA-2 family by Meta and MPT Models by MosaicML demonstrate the importance of scalability and transparency in LLM development [53]. By offering a range of model sizes and disclosing detailed information about their training data, these models enable developers to choose the most appropriate solution for their specific use case, balancing performance, efficiency, and personalization requirements [53].

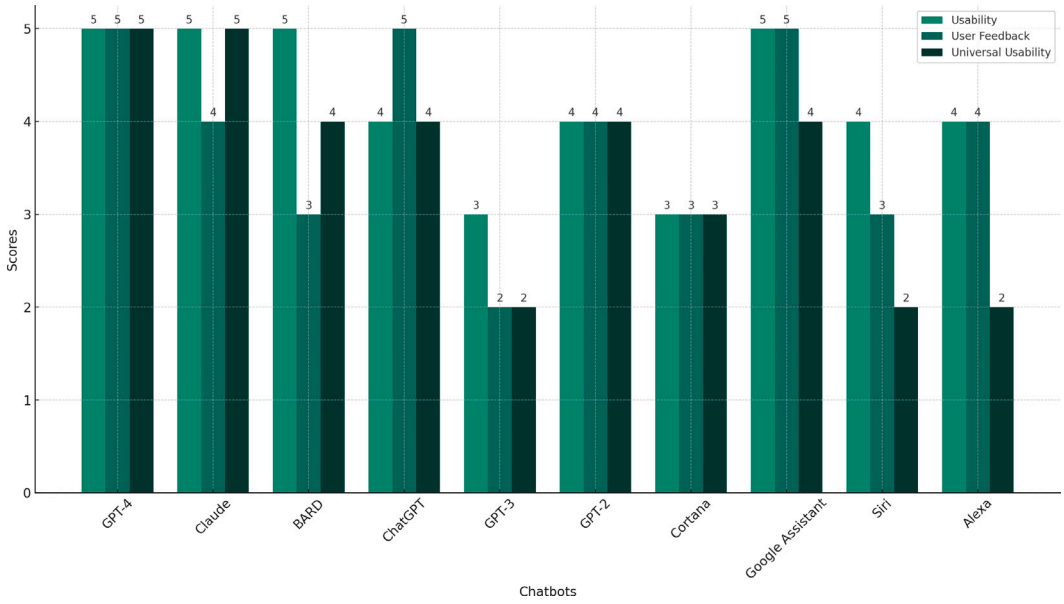

Fig. 11. Attributes of conversational AI platforms visualized by UI UX.

Table 5
LLMs and Personalization Capabilities.

LLM	Developer	Fine-Tuning capabilities	Personalization level
Gemini [45]	Google	Multimodal capabilities with state-of-the-art performance across various benchmarks, long-context processing up to 1 million tokens, and optimized versions for different tasks (Ultra, Pro, Nano). Fine-tuning for safety and content generation with detailed control over generation parameters and safety settings.	Very High: Offers granular control over content generation and safety, accommodating a wide range of user preferences and requirements. Its multimodal and long-context capabilities enable complex reasoning and understanding, providing personalized user interactions across diverse applications.
Claude 3 [46]	Anthropic	Customizable to suit a wide variety of needs with sophisticated vision capabilities, fewer refusals, improved accuracy, long context window, and responsible design for safety.	Very High: Prioritizes safety with a responsible AI framework, allowing for fine-tuned personalization within ethical and societal bounds. Its ability to handle complex, multi-step instructions enhances user experiences with personalized content that adheres to brand voice and guidelines.
GPT-4 [20]	OpenAI	Known for versatility and fine-tuning on a vast array of tasks with significant improvements in reasoning and understanding over GPT-3.5.	High: Provides adaptability and performance enhancement through fine-tuning and reinforcement learning from human feedback, making it suitable for a broad range of personalized applications.
LLaMA-2 [31]	Meta	Fine-tuning based on human preferences (RLHF), wide model range for different scales of tasks.	High: Supports customization and alignment with user preferences across different domains, though with potential limits in representing the full range of users' values and preferences.
MPT Models [47]	MosaicML	Commercial use licence, trained on 1T tokens with detailed training mix disclosure.	Moderate to High: Allows for specific task optimization though may have limitations in granular user-level personalization due to the broad scope of its training data.
Falcon Series [48]	TIIUAE	Detailed training data from diverse sources, large model variants with transparency in training processes.	High: Enables task-specific customizations with a focus on diverse data sources, allowing for a degree of personalized content generation.
StableLM Series [49]	StabilityAI	Fine-tuning with various datasets, including experimental and proprietary mixes, for extensive personalization.	High: Offers detailed fine-tuning capabilities for personalized outcomes, particularly in creative domains.
X-Gen [50]	Salesforce	Data scheduling system for fine-tuning, facilitating task-specific adaptations.	Moderate: Tailored for enterprise solutions with some level of personalization through task-specific data scheduling but may lack the depth of user-level customization.

The Falcon Series by TIIUAE and StableLM Series by StabilityAI showcase the value of diverse training data and fine-tuning capabilities in achieving personalized outcomes [53]. By leveraging a wide range of data sources and allowing for customization through fine-tuning, these models can generate content that closely aligns with user preferences and expectations [53].

Salesforce's X-Gen, while primarily focused on enterprise solutions, highlights the growing demand for LLMs that can adapt to specific tasks and industries [54]. As businesses increasingly rely on AI to improve their operations and customer experiences, models like X-Gen will play a crucial role in delivering personalized solutions that meet the unique needs of each organization.

Table 6
Well-known Evaluation benchmarks for LLMs.

Benchmark	Authors	Description	Performance summary
HumanEval	Chen et al. [55]	Evaluates the capability of language models to solve a diverse set of programming challenges by generating correct code solutions. This benchmark aims to measure the models' ability to understand and translate natural language descriptions into executable code.	Claude3 outperforms its counterparts with a score of 84.9%, while Gemini and GPT-4 followed with scores of 71% and 69%, respectively, and Llama2 trailed as the lowest performer with a score of 30.5%.
MMLU (Multitask Multilingual Language Understanding)	Hendrycks et al. [56]	Designed to evaluate models exclusively in zero-shot and few-shot settings. It ranges in difficulty from an elementary level to an advanced professional level, and it tests both world knowledge and problem solving ability. Ideal for pinpointing a model's blind spots.	Claude3 and GPT-4 emerged as the top performers in the evaluation, achieving scores of 86.80% and 86.40% respectively, while Llama2 lagged behind as the least effective model, with a score of 69.9%.
GSM8K (Groundedness, Situation Modelling, and Long-Range Reasoning)	Cobbe et al. [57]	Designed to test models' capability to understand and reason about realistic situations, requiring them to exhibit grounded situation modelling and long-range reasoning skills.	Claude3 has a leading score of 95%, closely trailed by GPT-4 and Gemini at 92% and 91% respectively, while Llama2 lagged significantly, recording the lowest score at 56.80%.
MATH	Hendrycks et al. [58]	Evaluates the model's understanding of mathematical concepts and their ability to translate natural language descriptions into mathematical operations.	Claude3 has the best score of 60.10%, with Gemini 1.5 Pro not far behind at 58.50% and GPT-4 close at 52.90%, while Llama 2 - 70B trailed significantly with a score of just 13.80%.
HellaSwag	Zellers et al.[59]	This benchmark measures the models' capability to comprehend visual information and use it to make inferences and predictions about real-world scenarios.	Claude3 achieved the highest score at 95.40%, followed by GPT-4 with 95.30% and Gemini 1.5 Pro at 92.50%; Llama 2 - 70B scored 87.00%, placing it last among the evaluated models.

Table 7
Model Comparison across prominent benchmarks [60].

Model	MMLU	HellaSwag	HumanEval	GSM8K	MATH
Claude 3 Opus	86.80%	95.40%	84.90%	95.00%	60.10%
GPT-4	86.40%	95.30%	67.00%	92.00%	52.90%
Gemini 1.5 Pro	81.90%	92.50%	71.90%	91.70%	58.50%
GPT-3.5	70.00%	85.50%	48.10%	57.10%	34.10%
Llama 2 - 70B	69.90%	87.00%	30.50%	56.80%	13.80%

3.6. Performance evaluation of large language models

Assessing the performance of chatbots and large language models (LLMs) is an important aspect of their development and deployment. It is crucial to consider a combination of different metrics and benchmarks to obtain a holistic understanding of their strengths and weaknesses. Various evaluation metrics and benchmarks have been proposed to measure different capabilities of these systems, ranging from natural language understanding and reasoning to task-specific skills like programming and mathematical problem-solving. This literature review provides an overview of some widely used evaluation metrics and benchmarks for chatbots and LLMs with a comparison of the most prominent models in Table 6

The evaluation of chatbots and large language models on benchmarks like HumanEval, MMLU, GSM8K, MATH, and HELLASWAG has shed light on the strengths and limitations of these systems. Recent LLMs such as GPT-4 (OpenAI), Claude (Anthropic) Gemini (Google), and LLaMA (Meta) have demonstrated promising performance on these benchmarks, outperforming earlier models like GPT-3.

From Table 7, it is observed that in MMLU benchmark, Claude achieved best performance at 86.80% with GPT-4 trailing closely at 86.40%. In terms of grounded situation modelling and long-range reasoning, as evaluated by GSM8K, Claude, GPT-4, and Gemini models outperformed GPT-3.5 and Llama2, with scores of 95%, 92.0%, 91.7%, 57.10%, and 56.80% respectively. When tested on challenging competition mathematics problems through the MATH benchmark, Claude clearly outperforms the rest with a score of 60.10% while Llama2 has the lowest score with 13.80%.

However, there is still room for improvement, as no single model has achieved outstanding performance across all benchmarks. Moreover, these benchmarks cover only a subset of the capabilities required for truly intelligent and versatile chatbots and LLMs. As the field continues to progress, new and more comprehensive evaluation metrics and benchmarks will be crucial for assessing the ever-evolving capabilities of these systems.

3.7. Regulations and compliance

AI-infused chatbots, such as ChatGPT, have the capacity to amass a significant amount of data from customer interactions, some of which could be classified as sensitive. It is crucial for businesses to handle this data in a manner that aligns with data protection regulations [61]. As AI regulations continue to evolve globally, they could potentially influence the adoption of AI tools like ChatGPT. For instance, the European Union is contemplating the classification of generative AI tools, including ChatGPT, as "high risk" in its forthcoming AI bill. This could result in these tools being subjected to stringent compliance requirements [62]. In response to these potential changes, many companies are already formulating policies around the use of AI tools. At the same time, experts are investigating the ethical implications of AI.

Sebastian [63] offers a comprehensive examination of the measures that can be implemented to protect user data in AI chatbots. This research delves into the dual tasks of preserving sensitive user data while maintaining the effectiveness of machine learning models. It scrutinizes existing Privacy-Enhancing Technologies (PETs) and suggests novel approaches, such as differential privacy, federated learning, and data minimization techniques. The research also incorporates a survey of Chatbot users to gauge their data privacy concerns associated with the use of these LLM-based applications.

In another noteworthy study, Arman and Lamiya [61] undertakes a crucial investigation into the ethical dilemmas associated with AI chatbots. The paper specifically zeroes in on the potential exploitation of AI chatbots in the dissemination of disinformation, a major issue in today's digital era. The study probes into potential regulatory measures that could mitigate these problems. In the fast-paced world of AI technology, the establishment of a sturdy regulatory framework that harmonizes the advantages of AI chatbots with the prevention of their misuse is of utmost importance.

Accuracy is an important factor when it comes to chatbots. Chatbots are generally more accurate than ChatGPT, as they are programmed to provide specific responses to specific inputs. ChatGPT and other language model-based chatbots may require more time and effort to set

up and maintain but can generate more sophisticated and human-like responses [64]. In short, while employing ChatGPT-powered chatbots, organizations must ensure that they are managing consumer data in accordance with data protection standards. Around the world, emerging AI policies may govern ChatGPT deployment. Companies should draft policies around the use of AI tools and experts should explore the ethical implications of AI. While chatbots are generally more accurate than ChatGPT, ChatGPT and other language model-based chatbots may require more time and effort to set up and maintain but can generate more sophisticated and human-like responses [30].

3.8. Fairness in large language models

Large Language Models (LLMs) have made significant strides in natural language processing, but their widespread adoption has also raised concerns about fairness and bias. This review examines the current state of research on fairness in LLMs, focusing on the nature and evaluation of bias, detection and mitigation strategies, and broader implications.

Gallegos et al. [65] propose a taxonomy of social biases in LLMs, categorizing them into embedding-based, probability-based, and generated text biases. Embedding biases arise from learned word representations, probability biases from predicted likelihoods favouring certain demographics, and generated text biases from the model's output perpetuating stereotypes. Li et al. [66] further distinguish between intrinsic and extrinsic biases. Intrinsic biases are inherent to the model's representations, while extrinsic biases manifest in downstream applications. This highlights the need for evaluating and mitigating biases at both levels.

However, detecting biases in LLMs is challenging. Husse and Spitz [67] show that bias detection methods are sensitive to minor variations in methodology. This underscores the need for robust and standardized evaluation frameworks that can reliably quantify biases across different models and tasks. Sun et al. [68] focus on gender bias in NLP. They propose a comprehensive taxonomy of gender bias mitigation strategies, spanning pre-processing, in-processing, and post-processing techniques. Pre-processing methods focus on curating and augmenting training data to reduce gender imbalances and stereotypes. In-processing techniques involve modifying the model architecture or training objective to promote fairness. Post-processing methods aim to detect and correct biased outputs after the model has generated them.

Weidinger et al. [69] broaden the scope of the discussion by considering the ethical and social risks posed by LLMs beyond fairness concerns. They outline six key risk areas, including discrimination, misinformation, malicious use, and environmental harms. This work underscores the need for a comprehensive risk assessment framework that considers the complex interplay between different types of harms and the potential trade-offs involved in mitigating them. For instance, efforts to reduce toxic language generation may inadvertently lead to decreased performance for certain demographic groups, highlighting the challenges of balancing competing objectives.

The importance of domain-specific considerations in understanding and mitigating bias in LLMs is highlighted by recent studies. Yuan [70] investigates the manifestation of bias in visual question answering systems, which often rely on LLMs for natural language understanding. The study reveals that these systems can perpetuate gender and racial stereotypes when generating answers to questions about images. For instance, the models may disproportionately associate certain professions or attributes with specific genders or races, even when the visual evidence is ambiguous or contradictory. This underscores the need for tailored evaluation and mitigation strategies that consider the unique challenges posed by multimodal AI systems.

In the context of professional language use, Thakur [71] examines how LLMs can introduce bias in job-related texts, such as resumes, job descriptions, and performance evaluations. The study finds that LLMs can amplify gender and racial biases in these texts, potentially leading to discriminatory hiring or promotion practices. Thakur [71] also highlights the subtle ways in which bias can manifest, such as through the use of gendered language or the association of certain soft skills with particular demographics. Mitigating these biases requires a nuanced understanding of the domain-specific language patterns and social dynamics at play.

These findings underscore the need for future research to prioritize the development of robust, domain-specific bias evaluation metrics. While general-purpose metrics can provide a high-level assessment of bias in LLMs, they may fail to capture the nuanced ways in which bias manifests in specific applications. By designing metrics that are attuned to the particular challenges and characteristics of different domains, researchers can more effectively identify and quantify bias in LLMs.

This section provided a comprehensive exploration of chatbots and AI-powered virtual conversational agents, with a specific focus on OpenAI's ChatGPT. It establishes an understanding of their functionalities, categorizations, and operational elements. Particular emphasis is given to the capabilities and limitations of ChatGPT, highlighting its ability to provide advanced responses while recognizing its need for additional features from a conversational AI platform to function effectively. The categorization of chatbots into task-oriented, social, knowledge-based, and hybrid types provides a systematic framework for assessing their diverse roles and functionalities.

Moreover, the review of existing AI-powered conversational chatbot frameworks presents a snapshot of the array of platforms in this sphere, underscoring the importance of these technologies in shaping the digital landscape. Crucially, the section accentuates the importance of datasets in training AI agents, UI/UX design in user engagement, and the need for data protection and compliance with emerging AI regulations. These insights drawn from the review are instrumental in guiding future research and development in the dynamic field of AI-powered conversation technology.

Building upon the understanding of AI-powered conversational chatbot frameworks, we will now turn our attention to their practical applications. These chatbot platforms have found use across a wide array of domains, each presenting unique requirements and challenges. In the following section, we explore these applications in detail, highlighting the transformative impact of chatbots in various contexts, from customer service to healthcare, and from education to entertainment. This exploration will provide a comprehensive view of their real-world utility and potential.

4. Applications

4.1. Education, learning and training

Kasneci et al. [72] gives a comprehensive overlook on the merits and demerits of using large language models such as Transformers in the education sector. The authors note the many applications of Transformers, such as educational content creation, including the personalization of learning experience which helps student engagement and interaction with the material. However, they bring with them their own obstacles. To name a few, as with any trained deep learning model, there is some bias involved in its output that is dependant on the data it is being trained. Moreover, the reliability of such models are still in question, thus demanding continuous supervision and oversight by humans. To address these challenges, Kasneci et al. [72] recommends developing competencies and documentations necessary to understand the nuances when using these models, especially when it comes to its shortcomings, and the way the technology should be used. Kasneci et al. [72] also suggested green technology based practices with efficient hardware and shared renewable energy powered infrastructures, regulatory compliance on data handling that insures ethical guidelines set out by a governing body, and upholding transparency to help remove biases. Policies, procedures, and other forms of government frameworks and controls to regulate the appropriate use of such models in the

education sector is also another important part in adopting such models in the education system. How do these deliberations translate to an application made for the educational domain in terms of performance is a question followed by Tack and Piech [73], Rodriguez-Torrealba et al. [74] and Sulaiman and Roy [75] in their paper.

Tack and Piech [73] reports on a first attempt at evaluating conversational models like Blender and GPT3 on its pedagogical abilities. Evaluation is done as a comparative measure of large language models against each other and a human teacher. The authors put forward a method to evaluate pedagogical ability of large language models by using a probabilistic model alongside Bayesian sampling. The method involves running these large language models built for conversation in parallel with humans with real world dialogue situations to test pedagogical abilities by comparing their responses for three abilities. The three abilities are speaking like a teacher, understanding students' needs, and helping them learn. Tack and Piech [73] made use of the Bradley–Terry model. Evaluation was conducted by an online survey, and the questions were chosen randomly from a set of items. The items contained a dialogue box, with a hypothetical conversation, replies from two random teachers of the three teachers considered, and three questions targeting pedagogical abilities. Tack and Piech [73] finally come to a conclusion that while the conversational agents were able to hold a conversation quite well, measured as conversational uptake, they were still significantly behind the abilities of human teachers, especially when it comes to helpfulness.

Rodriguez-Torrealba et al. [74] focused on generating questionnaires that consisted of multiple choice questions using large language models. Question generation was done based on Wikipedia articles as inputs. Rodriguez-Torrealba et al. [74] broke down the problem statement into three parts, namely Question Answering, Question Generation, and Distractor Generation. The authors used pre-trained T5 models for their application, and made use of the dataset DG-RACE. The questionnaires generated in such a manner were examined by experts, who all reported on the questions and the multiple options for answers given were mostly formed well, however, they seemed to favour knowledge retention more, than actual concept comprehension. Rodriguez-Torrealba et al. [74] have noted the absence of an adequate comparison metric for Distractor Generation and have put forward cosine similarity, that in this case would use word embeddings. The paper also discussed the limitations of the study, such as using only english Wikipedia articles and the need for further evaluation of the generated questionnaires. Overall, the paper presented a promising approach to assisting educators in generating assessments at a large scale, which would aid a rapidly growing education sector immensely.

Sulaiman and Roy [75] set out to address and review the problem of bias in automated decision making which ML models often showed. To this extent, they investigated the fairness of transformer models on two datasets, namely Student-Mathematics and Law School. Like previously noted concerns by Kasneci et al. [72] and Rodriguez-Torrealba et al. [74], the bias involved in the machine learning model used could prove a big deterrent in any application that would use such models. Therefore, to study effects of such occurrences Sulaiman and Roy [75] considered two models, one trained without bias mitigation method and one trained with bias mitigation methods. Sulaiman and Roy [75] used different fairness metrics to evaluate the trade-off between fairness and accuracy of models. The experiment was carried out for both datasets. In their paper, the authors note the SAINT model achieved optimum group fairness without requiring any debiasing methods. While, another model namely the Tab model, showed negligible decrease in accuracy with the use of debiasing methods. Empirically, the approach by [75] showed impressive results, and provided a well structured report regarding the trade-off between fairness and performance for transformer based models.

Another usecase of ChatBots is linguistic assistance or students. To this effect, Potier Watkins et al. [76] sets out to use transformer models

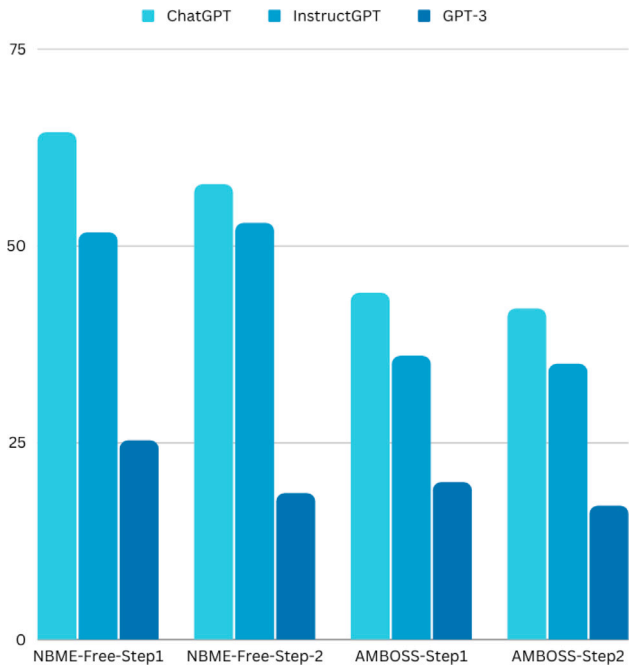

Fig. 12. Performance in terms of percentage of correct answers by models on different datasets [77].

to help resolve the difficulty in discovering the relevant grapheme-phoneme correspondences for effective phonics instruction, which usually requires scrutiny by a linguistic expert. They makes use of a software module called GPA4.0 that uses a Transformer model to automate the task of transcription of graphemes to phonemes. The module is trained on four different languages, which sequentially decrease in orthogonical transparency. The languages are spanish, portuguese, french, and english. The results by Potier Watkins et al. [76] shows that the Transformer model used in GPA4.0 show significant improvement over current state of the art models. The results mainly attributed to the attention mechanism of transformers allowing better alignment between graphemes and phonemes. Overall, Potier Watkins et al. [76] demonstrate a practical approach on developing and assisting educational material with the use of Transformer models.

Gilson et al. [77] set out to evaluate the academic capabilities of the Language models. Their research evaluated the performance of ChatGPT on the United States Medical Licencing Examination Step 1 and Step 2 exams. Their study used two sets of multiple-choice questions, one derived from AMBOSS, a medical learning knowledge resource, and the other from the National Board of Medical Examiners, NBME's free questions. A comparison was done with ChatGpt with two other large language modles. The models were InstructGPT and GPT-3, which is the underlying architecture of ChatGPT. The results by Gilson et al. [77] showed that ChatGPT showed around a 8.15% increase in performance than InstructGPT, while GPT-3 performed akin to random chance. ChatGPT achieved accuracies of 44% to 64.4% across the datasets reviewed. The models were judged on different aspects, including justification given to the answer, presence of information both internal and external to the question. While the model's performance dipped noticeably the harder the questions got, performing at a greater than 60% threshold on the NBME-Free-Step-1 data set, the model achieved a promising level of accuracy as can be seen in Fig. 12. The paper overall gave a concise review on the current capabilities of ChatBot and their supporting technology when tested on academic questions.

4.2. Business and customer care

Transformer models are mainly used for language modelling and ChatBot applications, but can be adopted for multiple sequence modelling tasks. The reason one might be inclined to do so is for real life scenarios that occur frequently, such as in business. Bukhsh et al. [78] proposes a deep learning-based approach called ProcessTransformer for predictive business process monitoring(PBPM). PBPM is the task of using event logs and other data sources to predict future events and characteristics of a running process. The proposed model by Bukhsh et al. [78] establishes dependencies between a multiple of event sequences and their outputs. The architecture itself involves long range memory and using a self attention mechanism. The event logs are sequentially ordered as they occurred in the actual runtime of the process, and past events are used to monitor future performance of a running process. The model was reported to outperform several currently used techniques and obtained on average an accuracy of above 80%. Bukhsh et al. [78] puts forward that this approach can be used to optimize business processes and improve customer satisfaction, reduce costs and increase efficiency in various domains, automate decision making processes, improve resource allocation and utilization in various domains and improve the quality of services the business offers.

In many businesses, heavy paperwork often convolutes many details that become a chore to sort out. Douzon et al. [79] set out to improve LayoutLM, which is a pre-trained language model used for information extraction in business documents such as expense receipts, invoices, and purchase orders. The authors introduced two new pre-training tasks to improve the model's ability on identifying relevant information from these documents. The first task was aimed at making the model understand the layout of the document, while the second task handled numeric values, and their magnitudes. The result was the model forming better contextualizations of the documents that were being scanned by the model. The authors evaluated their method on both public and private datasets and showed significant improvements in extraction performance. The paper also discussed the evolution of attention-based document analysis models, such as Transformers and BERT, and how they have improved state-of-the-art performance. Overall, the results shown by Douzon et al. [79] showed significant improvements in performance over older models deployed.

Heidari and Rafatirad [80] proposed a transfer learning approach for a model, aimed at assisting safer investment decisions in the real estate market. The model used natural language processing techniques to analyse online textual information from real estate websites to detect the potential of a house as a rental property. The research used bidirectional attention models based on transformers extract features, similar to that used by BERT. This features where then put through a convolutional neural network for further processing. The model was trained on a new public dataset of more than 5 million houses in the US, and was evaluated against traditional machine learning models for rent prediction. The study applied weighted balance in the loss function to counter the fact of there was an imbalance of house types in each zip code. Heidari and Rafatirad [80] achieved a significant improvement over the other models involved in comparison, achieving upto a 10% improvement in f1 scores.

Looking at the Human Resources component of business, Agarwal et al. [81] evaluated different transformer models, their main focus being to experimentally conclude that transformer models trained and finetuned on factors such as emotions prove to be model that better mimics human response and has better performance than a model that has no such finetuning. The models being compared were two retrieval model, that retrieves a response from a corpus of response based on the prompt, and one generative model, that works like a language model generating a sequence of words based on the prompt. All models had a baseline version, as well as versions that were finetuned on the Empathetic Dialogues dataset, or the Twitter Customer Care dataset, or both. The two Retrieval models, one that used a bi-encoder, and one that used a poly-encoder, were evaluated using two baseline architectures described by Humeau et al. [82]. The generative model was evaluated using the perplexity model. Across all three models, a significant rise in performance was observed from the baseline models to the finetuned models. While finetuning on the Empathetic Dialogue database had non perceptible benefits on the generative model, it proved to increase the performances of both the retrieval models.

4.3. Public health and medicine

ChatBot implementation would prove to be a huge tool in Health Industries if set up and used properly. Yadav et al. [83] studied the task of summarization of frequently asked real-world consumer health questions into its abstract. The model proposed leveraging semantic interpretation of a question by identifying the medical terminologies and the specific entities of the context, from which the summaries are generated. Two models were looked into, a Question Focus aware summarization model that uses three Cloze tasks of sequence-to-sequence masking, N-gram masking and finally medical entity masking. The second was a Question Type aware summarization model that used question type information infused into decoder inputs, enabling generation of question-type driven summaries, which was done by a finetuned MINILM Transformer based pre-trained model used on MEQSUM dataset for summarization. These models were evaluated against baselines of pretrained transformer models such as T5, PEGASUS, BART, using standard ROGUE metrics. The proposed models displayed better scores then the baseline models, off of which the paper concluded knowledge of question types and awareness of salient medical focus words leads to more informative and relevant summarizations.

Li et al. [84] presented a deep neural sequence transduction model called BEHRT for predicting the likelihood of more than three hundred conditions in one's future visits using health records. BEHRT can predict multiple diagnoses simultaneously. The model was tested on CRPD, against the other state of the art models and showed a noticeable improvement in accuracy over current state of the art deep electronic health records models. Data was gathered from about 1.6 Million people to train the model. The relationship between diseases are studied using the attention mechanism of the transformer models. Multiple heterogeneous concepts can be considered and incorporated for better accuracies. The disease, patient representations provided by the results can prove extremely useful in any further studies conducted. The main objective of [84] was to build a well performing model for the future diseases prediction, by providing multiple by products of training the model that gives us useful insights. The embeddings resulting from the model could prove to be a great perspective into the interconnectivity of cases of different diseases.

Polignano et al. [85] looked at the task of classifying named entities in text into defined categories such as person, organization, and location, from medical documents using deep learning. The authors proposed a hybrid approach that combines transformer-based models and conditional random fields for analysing textual medical diagnoses. Polignano et al. [85] compared different transformer architectures such as BERT, RoBERTa, and ELECTRA to find the best performing model for the task. Li et al. [84] also highlighted the importance of automated analysis of medical documents and the difficulties encountered with short and specific documents, concluding that the effectiveness of predictions is crucial in this field to avoid mistakes that could compromise people's lives. Overall, the paper presented a hybrid approach for Named Entity Recognition(NER) from medical documents compared different transformer architectures to find the best performing model.

4.4. Other applications

Other uses of chatbots are limited in variety, most sizeable of which is in computer science applications such as generating code, database management and bug fixing.

Lin et al. [86] used ChatGPT to manage databases using Context-based Ontology Modelling for Databases (COM-DB), an approach for transforming database schema into plain language. In database operations, commands are referred to as queries. ChatGPT is yet to produce queries. This is due to database schemas by nature having critical information about database structures that are usually expressed in graph form rather than normal language, and there is no standard way to convey data semantics. COM-DB supports plain language syntax for representing data semantics such as table structure and relationships. COM-DB makes use of terms such as "context-of", "monodirectional relationship", and "bidirectional relationship". "context-of" is the topic of the investigation reported in the paper. Synthea-Alabama and BDA-EHR sample datasets were obtained and used. Two studies were carried out to illustrate common database integration tasks: Semantic Integration and Tables Joining. Semantic Integration involved the merging of two tables with the same kind of information, whereas Tables Joining entails creating a new table that incorporates data from many tables. The studies were carried out with and without COM-DB, and each experiment was repeated ten times to average out the fluctuations introduced by ChatGPT. Experiments based on COM-DB were more successful and accurate. Lin et al. [86] were led to assume then that ChatGPT can be used in the industry for database operations through COM-DB, since its benefits include quicker database administration, decreased domain knowledge, and greater privacy protection.

Song et al. [87] suggests developing a web-based chatbot for interactive BUg repoRTing (or BURT). Given the limits of existing bug reporting systems, the problem solved by BURT is assuring state-of-the-art bug reporting. BURT used natural language processing, dynamic software analysis, and automated bug report quality evaluation methodologies. BURT is made up of three primary parts. The Natural Language Parser parses bug descriptions provided by users. The discussion flows for the reporting process are implemented by the Dialogue Manager. The Report Processing Engine compares the parsed bug descriptions to the app execution model to determine the quality of bug elements and give recommendations. Twelve Android bugs were chosen from six different Android apps: AntennaPod, Time Tracker, GnuCash, GrowTracker, and Droid Weight. People were recruited to utilize BURT and report bugs. Their interactions with the chatbot were recorded and analysed. The recommendations were valuable to half of the reporters, and they were occasionally useful to the other half. As a result, employing chatbots as a way of bug reporting offers advantages that make it a realistic alternative in practice.

It is observed from Fig. 13 that the market revenue of chatbots have increased from 2018 till present, and is projected to further increase dramatically. The chatbot market is expanding rapidly, owing mostly to the increased usage of chatbots to manage enquiries and provide customer service. [1,6,7]. Increased speed of handling issues and better turnover rate means that chatbots are efficient tools for business. It is also seen that customers have become increasingly more aware of chatbots and accepting of them which makes the transition easier as time progresses [89]. Apart from the large market in USA, the market in Asia is also projected to increase rapidly in the coming years [89].

Throughout all papers reviewed, some of them had sections discussing the implications ChatBot and NLP applications and techniques might have on he industry, and the shortcomings of current technology that need to be addressed. The following section goes over the main ones identified throughout the review.

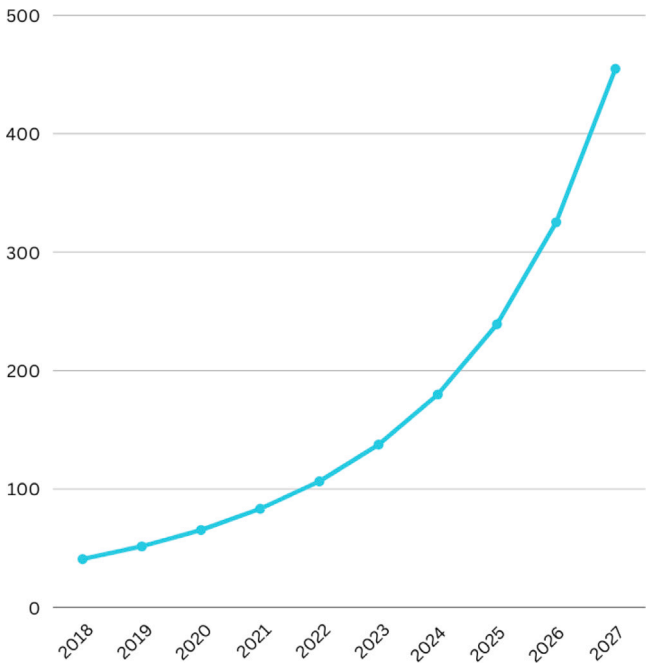

Fig. 13. Market revenue of chatbots by year (in million USD) [88].

5. Open challenges

The 4 main challenges (see Fig. 14) of ChatGPT and similar AI chatbots identified to be common among reviewed papers are of hallucination, biases in training data, jailbreaks, and anonymous data collection [90–93].

Hallucination is the phenomenon where AI chat bots like ChatGPT synthesize factually incorrect responses that otherwise appear viable [90]. The very nature of ChatGPT is to "hallucinate" as it is a sequence prediction model, and uses it is pre-trained weights to guess the next most likely sequence of words irrespective of whether it is factual or not. This is being tackled partially by the creator of ChatGPT OpenAI in their next release- GPT-4. The use of a Neural-Retrieval-in-the-Loop methodology of Retrieval-Augmented Generation (RAG) is one way being investigated to lessen hallucination [94] to combine parametric knowledge (knowledge that is stored within the weight values of the parameters that constitute the model) and non-parametric knowledge (knowledge that exists outside the model, such as search engines, or wikipedia) through a query generation process. This leads to state-of-the-art results in knowledge-intensive tasks and reducing hallucination. This can be improved upon by employing Poly-encoder Transformers and end-to-end-trained retrievers in the Fusion-in-Decoder technique [95].

Since chatbots are trained on natural language text from documents and websites on the internet, the models are susceptible to learn biases and stereotypes that may be present in the text that it is trained on. This may include racist, xenophobic and misogynistic language. Therefore the kind of biases that are trained into chatbots can cause and proliferate misinformation, harm and toxicity [90]. Even when using human data labellers who follow recommendations for training and fine-tuning datasets, often it is not helpful as data labellers are not representative of varied ideas and perspectives, which introduces biases into data [91].

When reinforcement learning through human feedback is utilized (like in ChatGPT), models are potentially prone to erroneous or harmful behaviour (e.g., issuing racist statements) [90]. This is due to the fact that humans are integrally engaged in a model's performance and knowledge, and when given fundamentally opposing inputs, values,

and preferences, its behaviour becomes chaotic. The lack of societal agreement on a single unified moral theory implies that there is no metric or norm with which the model may fix its responses [96]. Nonetheless, RLHF models have made progress in decreasing bias since they use human feedback to directly and efficiently address algorithmic prejudice in compared to prior methodologies [97]. It may also be used to counteract long-standing impacts of historical, representation, and measurement bias by balancing human feedback with representation and expertise among a varied spectrum of human annotators. [96].

[98] identifies that the use of "personas" by chatbots is another circumstance that frequently results in unpleasant responses. For instance, altering ChatGPT's system parameter and giving it a character, like boxer Muhammad Ali, greatly raises the generational toxicity. The toxicity of ChatGPT can rise multitudes based on the persona assigned to it, with outputs participating in false stereotypes, unpleasant conversation, and nasty viewpoints. Regardless of the persona assigned, certain entities are targeted more frequently than others, which reflects the model's intrinsic discriminatory biases.

The problem of bias can be solved by vetting the resources used to train models, training them on a larger volume of data as well as hard coding instructions and responses to certain prompts. One method as proposed by [99], of vetting and de-biasing data by using a text style transfer model. By combining key components of Explicit Style Keyword Replacement models and Auto-encoder sequence-to-sequence models, the resultant model can convert a piece of biased text to a neutral version whilst maintaining significant content information.

Jailbreaks refer to instances where AI chatbots may deviate from their intended tasks or constraints and produce undesired responses. For example, ChatGPT might occasionally give a response that is inappropriate or irrelevant, despite the safeguards put in place [90,92,100,101].

[101] speaks about how cybercriminals can employ LLMs to carry out cyberattacks by either directly obtaining the information or getting beyond Open-AI's moral guidelines. Attackers generate convincing social engineering attacks, phishing attacks, attack payloads, and various types of malicious code snippets that can be assembled into executable malware files. Although ChatGPT and other LLMs are prohibited by OpenAI's ethical guideline from explicitly providing malicious information to attackers, there are ways to get around the limitations placed on these models using techniques like reverse psychology.

Jailbreaks are caused mainly due to the complexity of natural language and therefore the sheer combinations of potential user inputs. A variety of taxonomies and classifications for types of jailbreaks have been hypothesized to help solve this issue, and a wider understanding of jailbreaks can be attained from knowing these classifications. One such attempt [92] classifies them into "Instruction-based jailbreak transformations" and "Non-instruction based jailbreak transformations". The first is further divided into three types. First is "Direct Instructions" where a model is told to ignore it is previous prompt and a new task is specified. Second is "Cognitive Hacking" where a model is provided a safe space or a situation that warrants an illicit response. Lastly, is "Instruction Repetition" where the same instruction is fed multiple times until the illicit response is obtained. Non-instruction based jailbreak transformations are also divided into three types. First is "Syntactical Transformation" uses orthographic transformation to bypass content filters. Second is "Few-shot Hacking", where the attack prompt contains several examples of text that may be designed to maliciously misalign the model. Lastly, is "Text Completion as Instruction", where attacks are incomplete prompts, and the model is forced to complete the sentence in a way that is illicit. Other classification attempts like that of [102] offer similar insight into the variety and effectiveness of jailbreaking prompts.

Security against jailbreaks is then a natural and existential concern for chatbots and their creators. To mitigate this, stricter guidelines are put in place, and continuous feedback loops are used to improve the model's performance. [102] has found that OpenAI puts tight limitations on subjects including compromising user privacy, illegal conduct, dangerous content, and fraudulent deceptive practices. In those circumstances, ChatGPT only returns the forbidden content rarely. Repeatedly inputting prompts also increases chances of jailbreaks, which means it also may not be strong enough to withstand continuous conversation. Attackers successfully got over prohibitions in cases of political campaigning, lobbying, and government decision-making. This shows that although these instances are on OpenAI's ban list, there do not appear to be any limitations in place, which raises questions about how easy it would be to access things that is prohibited.

New methods to train and prepare models for jailbreaks have been considered. [100] argues against the notion that scaling alone can solve these safety failure modes and emphasizes the necessity for safety-capability parity, which states that safety procedures should be as complex as the underlying model. [100] hypothesizes safety training methods based on two failure modes— "competing objectives" and "mismatched generalization". Competing objectives is when a model's safety aim conflicts with its pretraining and instruction-following objectives, and mismatched generalization is when inputs are outside of distribution for a model's safety training data but within the bounds of its extensive pretraining corpus. [100] contends that jailbreaks might be a natural result of the safety teaching techniques already in use. Scaling up will not address conflicting goals because the root cause of this failure mode is the optimization objective rather than the dataset or model size, in fact. If safety training is not properly extended to further domains, it can make mismatched generalization worse. The ability of safety training to generalize as broadly as the capabilities of models cannot be guaranteed by more data or larger models. Scaling may instead expand the attack surface.

Once again, a special case of jailbreaks is when the model was trained with RLHF. [96] discusses that because of the nature of learning through human feedback, a model can over fit to a style of maximizing the chance that the output answers are most likely to be seen as viable to the user. This leads to increased susceptibility to jailbreaks when these models are put into production as these models try to fulfil the prompts in the way that the user wants it to. However stricter RLHF methods can be used to get stronger rules, that can mitigate these harms, and the methods used for mitigation of bias are also applicable here.

In Context, learning is the process through which transformer-based LLM models learn from instances in the context. In it, we give the LLM a cue in the form of a sequence of input–output pairings that demonstrate a task during in-context learning. To get a prediction, we insert a test input to the end of the question and then let the LM condition on the prompt and forecast the remaining tokens. In order to respond effectively to the prompts, the model must read the training samples to determine the input distribution, output distribution, input–output mapping, and formatting. [103] proposes a three step approach to use in-context learning to avoid undesirable responses to prompts. "Safety demonstrations" are dialogue data that represent safe model behaviour. These are used to condition a model's generation of output. The responses are ranked based upon their similarity to said safety demonstrations.

Lastly, the challenge of anonymous data collection has to do with the ethical use of data. AI models like ChatGPT are trained on very large amounts of data, and while they are designed not to remember specific interactions, there may still be concerns about privacy and data security [91]. Due to the nature of anonymous data collection, these LLM chatbots could unintentionally leak private user information. Furthermore, many of the databases that ChatGPT may access are obtained from the Internet, including social media sites such as Twitter. As a result, in addition to privacy issues, ChatGPT may identify content that threatens personal privacy. Another concern related to privacy is that individuals may also be tracked and profiled based on their ChatGPT interaction histories [91]. The challenge here is to ensure

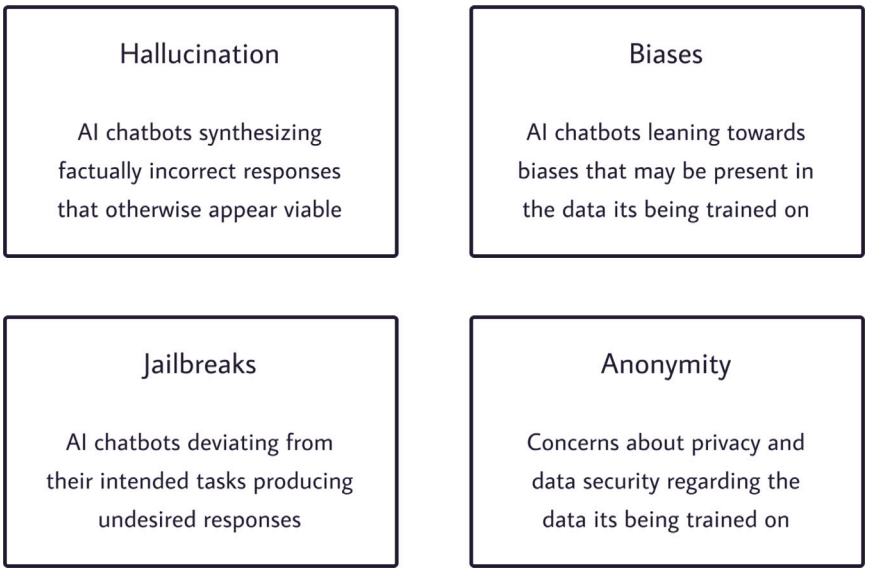

Fig. 14. The four main limitations that exist for LLMs.

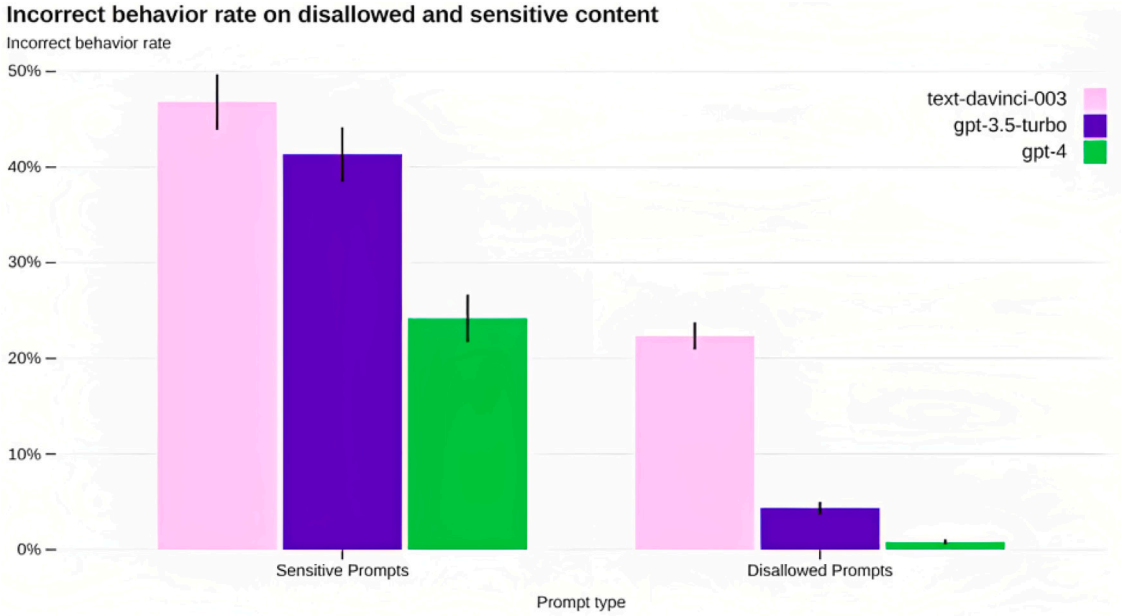

Fig. 15. Jailbreak tendency of different models [51].

that the AI maintains the privacy of its users while still improving its understanding and performance. OpenAI is addressing this issue through strict data usage policies, ensuring no personally identifiable information is stored or used in the training of the model. Furthermore, OpenAI is currently actively researching methods for differential privacy, which will allow the model to learn from data without risking the privacy of individual data points.

As mentioned before, training LLMs that are used to make these Chatbots raise a lot of ethical concerns that need to be considered and addressed. LLMs can amplify a lot of biases in training data from different sources (see Fig. 16) that is reflected in the final result, as well as hallucinations that generate fake information [91,104,105]. [104,105] also raise the problem of cultural homogenization, that is LLMs inadvertently marginalizing minority voices and local nuances, leading to the diminishing of cultural diversity. Cultural nuances and significance can be lost during training of LLMs due to under representation, and meanings being lost during translation. LLMs lack transparency [91,105], making it difficult to understand, justify or rectify the decisions they make. [91] also expands on ethical concerns that surrounds authorship of papers or any other content. [106] postulates LLMs, during training, can assign "mental states" to individuals with the data they train on, but cannot empathize like humans do. LLMs' method of assessment is based on predicting language tokens to generate a response that fits the training data, rather than considering how an individual may be an outlier. Empathy is a way to respect the individuality of a person or the situation, which the accuracy that an LLM revolves around cannot account for [106].

To make LLMs more inclusive of ethical considerations, frameworks and policies should be made to integrate them during the training and usage of LLMs. [91] makes recommendations on usage of LLMs based on the ethical concerns listed specifically made for different stakeholders identified by the paper, such as researches, developers, users, social scientists etc. [107], while noting the bias towards western and english speaking societies in GPTx models, proposes the need for 'in-context'

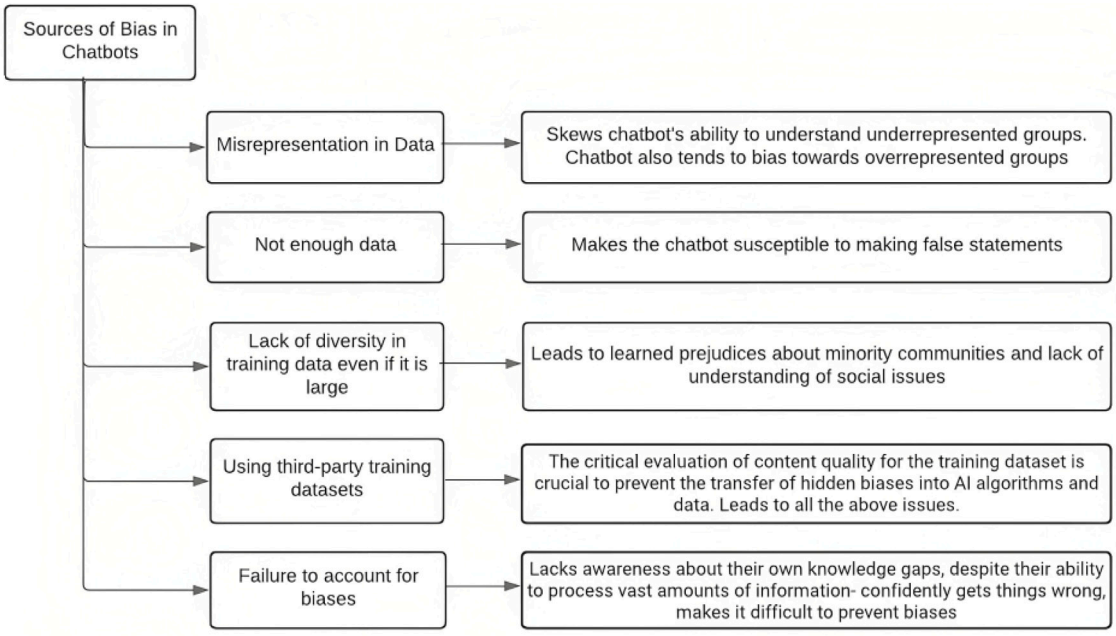

Fig. 16. Sources of bias.

ethical policies based on the different branches of normative ethics such as virtue, deonotlogy and consequentialism. These policies are designed to be specific to the cultural, linguistic, and ethical context in which the LLM operates. The framework allows for the specification of variables and values on which the policies apply, ensuring that the ethical principles are applied appropriately in different situations. [108] looks specifically into the use of AI in medical fields. As noted by [105], ethical concerns have more dire concerns in sensitive fields such as medicine. To address that, [108] proposes ten commandments to help mitigate the issues of using AI in medicine. The commandments seek to make the use of AI transparent, comprehensible and repeatable, carried out by competent individuals based on state of the art theories.

The OpenAI research article [51] announcing GPT-4 goes through the performance of GPT-4 in comparison with previous models with various metrics. In Fig. 15, we observe that GPT-4's performance is far better than the others. OpenAI attributes this to the rigorous care undertaken by them in order to ensure that a model as powerful as GPT-4 is not misused. Experts from multiple different AI alignment related fields studied GPT-4's behaviour in high-risk regions. These experts' feedback and data increased the model's capacity to decline risky requests. During RLHF training, the model additionally adds a safety incentive signal to limit dangerous outputs. Furthermore, a varied dataset is gathered to prevent the model from rejecting genuine requests. Keeping some of these obstacles in mind, the next section discusses possible future work that may be done to solve some of these concerns, as well as additional areas for improvement.

6. Future research directions

As the field of AI chatbot applications using large language models continues to evolve, there are several areas that require further exploration and development. Some potential future directions and areas of focus for research and improvement in this domain have been recognized by us and the reviewed papers. There are two main distinct fields, or perspectives on improvements that need to be made on research surrounding Chatbots (see Fig. 17). For technical issues, one of the challenges identified in our last section with current LLMs was their language bias. As many of our reviewed papers noted, most LLMs are trained on a majority of English language data, which can result in limitations when applied to other languages. Additionally, even in multilingual LLMs, there is often a skew towards the more common language types, leaving many languages underrepresented in the training data. To overcome these limitations, future work should focus on modifying LLMs to mitigate language bias. One potential approach is to incorporate a mixture of both monolingual and multilingual LLMs during training. By doing so, LLMs can be exposed to a more diverse range of languages, enabling them to better understand and generate content in multiple languages. Moreover, efforts should be made to include more representative and balanced training data for underrepresented languages, ensuring that LLMs are more inclusive and effective in supporting diverse linguistic needs. While LLMs have demonstrated impressive language generation capabilities, there is room for improvement in terms of user interaction and personalization. Future work can also focus on developing chatbot interfaces that facilitate more natural and intuitive interactions with users. This could involve integrating multimodal inputs, such as voice and gestures, to enable more dynamic and engaging conversations. Additionally, personalized adaptation of LLMs should be explored to cater to individual user preferences, learning styles, and knowledge levels. By incorporating user feedback mechanisms and adaptive algorithms, LLMs can provide more tailored and contextually relevant responses, enhancing the overall user experience and effectiveness of AI chatbot applications in education. On the other perspective that requires dressing ethical concerns of ChatBots, LLMs have the potential to greatly impact decision-making processes in various domains. However, concerns have been raised about the ethical implications of LLMs making critical decisions. For example, in the Educational Domain, a use case might be to use such models for providing academic advice or grading assignments. Future work should focus on evaluating further and enhancing the ethical decision-making capabilities of LLMs. This can be achieved by incorporating ethical guidelines and principles into the training process, enabling LLMs to align their responses and actions with ethical standards. Additionally, research should explore methods to provide explanations and justifications for the decisions made by LLMs, allowing users to understand the reasoning behind their suggestions or assessments. By prioritizing ethical considerations and transparency, LLMs can be used in a manner that promotes fairness, trust, and accountability in educational settings. Taking the legal viewpoint and to ensure the appropriate use of LLMs in any sensitive domain areas, such as education and healthcare, future work should emphasize the development

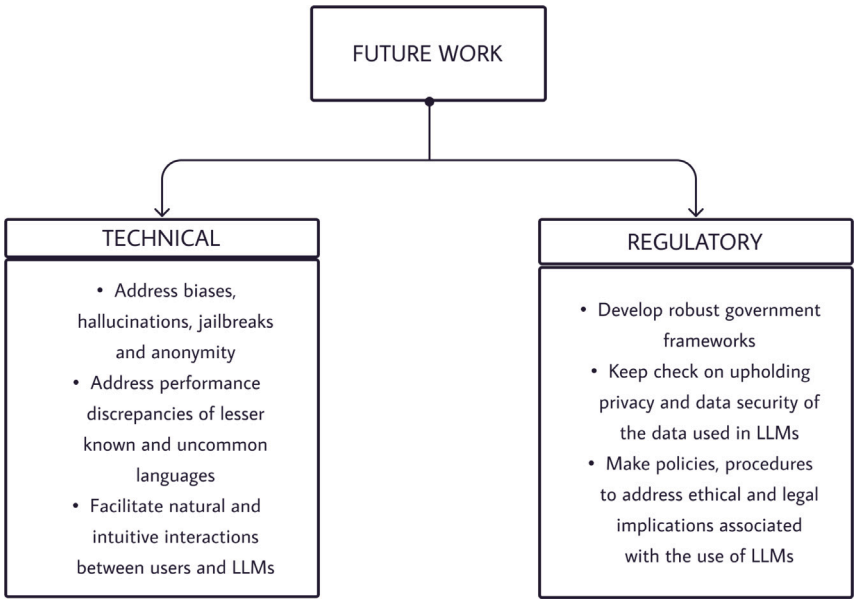

Fig. 17. Two different perspectives of areas of improvement in LLM Technology Development.

of robust governance frameworks. These frameworks would encompass policies, procedures, and controls aimed at addressing the ethical and legal implications associated with the use of LLMs. As LLMs become more prevalent in these settings, it becomes imperative to establish guidelines that govern their deployment and use. These frameworks should take into account factors such as data privacy, algorithmic bias, accountability, transparency, fairness, and reliability. Researchers and policymakers should collaborate to design and implement governance frameworks that strike a balance between innovation and responsible use of LLMs.

Another shortcoming surrounding current research done around ChatBots is the lack of standard metrics of comparison that is common between papers to help compare new emerging technology on various relevant aspects of text generation tasks, such as reliability and accuracy of information, readability and other conversational qualities of the application. A standard baseline with measurable qualities that represent ethicality and compliance to guidelines and regulations could also prove crucial to help streamline most applications to desired levels of ethical security.

7. Conclusions

This review has provided a thorough examination of the current state of AI-powered virtual conversational agents. An in-depth comparison of frameworks, datasets and UI/UX was conducted and its findings were reported. We have then explored its diverse applications across various sectors, including education, business, public health, and more, demonstrating its potential to bring about significant transformations. However, we have also highlighted the challenges that these AI chatbots face, such as hallucination, biases in training data, jailbreaks, and anonymous data collection. These challenges underscore the need for ongoing research and development to ensure the responsible and effective use of these tools. Looking to the future, we have identified several key research directions that hold promise for enhancing the capabilities of AI chatbots. These include efforts to mitigate language bias, enhance ethical decision-making capabilities, improve user interaction and personalization, and develop robust governance frameworks. By addressing these areas, we can ensure that AI chatbots like ChatGPT can be used effectively and responsibly across a wide range of applications. AI chatbots like ChatGPT represent a significant advancement in the field of AI and natural language processing. They hold immense potential to revolutionize various sectors, from education and business to public health and beyond. With continued research and development, these tools can help overcome many of the current challenges faced in several domains and unlock new possibilities for the future.

CRediT authorship contribution statement

Avyay Casheekar: Conceptualization of this study, Methodology, Categorization. **Archit Lahiri:** Expanded the scope of applications of chatbots, Created visualizations across sections, Detailed on the open challenges and future works. **Kanishk Rath:** Handled selection of papers, filtering and selection criteria. Reviewed a portion of the papers and created visualisations for the same. **Kaushik Sanjay Prabhakar:** Handled selection and filtering of prior review papers on the topic. Reviewed a portion of the papers and identified the shortcomings in the literature. **Kathiravan Srinivasan:** Reviewed the organization, Guided the review.

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.

Data availability

Data will be made available on request.

References

[1] G. Caldarini, S. Jaf, K. McGarry, A literature survey of recent advances in chatbots, Information 13 (1) (2022) http://dx.doi.org/10.3390/info13010041.
[2] S. Singh, H. Beniwal, A survey on near-human conversational agents, J. King Saud Univ. - Comput. Inf. Sci. 34 (10, Part A) (2022) 8852–8866, http://dx.doi.org/10.1016/j.jksuci.2021.10.013.
[3] A. Al, S. Singh, A.B. Kocaballi, M. Prasad, An overview of conversational agent: Applications, challenges and future directions, 2021, pp. 388–396, http://dx.doi.org/10.5220/0010708600003058.
[4] S. Hussain, O. Ameri Sianaki, N. Ababneh, A survey on conversational agents/chatbots classification and design techniques, in: L. Barolli, M. Takizawa, F. Xhafa, T. Enokido (Eds.), Web, Artificial Intelligence and Network Applications, Springer International Publishing, Cham, 2019, pp. 946–956.

[5] D.-M. Park, S.-S. Jeong, Y.-S. Seo, Systematic review on chatbot techniques and applications, J. Inf. Process. Syst. 18 (2022) 26–47.

[6] G. Bilquise, S. Ibrahim, K. Shaalan, Emotionally intelligent chatbots: A systematic literature review, in: Z. Yan (Ed.), Hum. Behav. Emerg. Technol. 2022 (2022) 1–23, http://dx.doi.org/10.1155/2022/9601630.

[7] E. Adamopoulou, L. Moussiades, Chatbots: History, technology, and applications, Mach. Learn. Appl. 2 (2020) 100006, http://dx.doi.org/10.1016/j.mlwa.2020.100006.

[8] E.H. Almansor, F.K. Hussain, Survey on intelligent chatbots: State-of-the-art and future research directions, in: L. Barolli, F.K. Hussain, M. Ikeda (Eds.), Complex, Intelligent, and Software Intensive Systems, Springer International Publishing, Cham, 2020, pp. 534–543.

[9] E. Adamopoulou, L. Moussiades, An overview of chatbot technology, ISBN: 978-3-030-49185-7, 2020, pp. 373–383, http://dx.doi.org/10.1007/978-3-030-49186-4_31.

[10] J. Weizenbaum, ELIZA—A computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1) (1966) 36–45, http://dx.doi.org/10.1145/365153.365168.

[11] H. Shah, A.L.I.C.E.: An ACE in digitaland, 4, 2005.

[12] ChatGPT, 2022, https://openai.com/blog/chatgpt. (Accessed 13 July 2023).

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018.

[14] A. Bernardini, A. Sônego, E. Pozzebon, Chatbots: An analysis of the state of art of literature, 2018, p. 1, http://dx.doi.org/10.5753/wave.2018.1.

[15] C.-C. Lin, A. Huang, S. Yang, A review of AI-driven conversational chatbots implementation methodologies and challenges (1999–2022), Sustainability 15 (2023) 4012, http://dx.doi.org/10.3390/su15054012.

[16] B. Galitsky, Adjusting chatbot conversation to user personality and mood, ISBN: 978-3-030-61640-3, 2021, pp. 93–127, http://dx.doi.org/10.1007/978-3-030-61641-0_3.

[17] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models, 2023, arXiv:2304.01852.

[18] T. Sakirin, R.B. Said, User preferences for ChatGPT-powered conversational interfaces versus traditional methods, Mesop. J. Comput. Sci. 2023 (2023) 24–31, http://dx.doi.org/10.58496/MJCSC/2023/006.

[19] D.-C. Toader, G. Boca, R. Toader, M. Măcelaru, C. Toader, D. Ighian, A.T. Rădulescu, The effect of social presence and chatbot errors on trust, Sustainability 12 (1) (2020) http://dx.doi.org/10.3390/su12010256.

[20] OpenAI, GPT-4 technical report, 2023, arXiv:2303.08774.

[21] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020, arXiv:2005.14165.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[23] Google bard, 2023, https://bard.google.com/. (Accessed 13 July 2023).

[24] Claude, 2023, https://www.anthropic.com/index/introducing-claude. (Accessed 13 July 2023).

[25] IBM watson, 2010, https://www.ibm.com/watson. (Accessed 13 July 2023).

[26] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, X. Huang, A comprehensive capability analysis of GPT-3 and GPT-3.5 series models, 2023, http://dx.doi.org/10.48550/arXiv.2303.10420, ArXiv, arXiv:2303.10420.

[27] J. Si, A.G. Barto, W.B. Powell, D. Wunsch, Reinforcement learning and its relationship to supervised learning, in: Handbook of Learning and Approximate Dynamic Programming, 2004, pp. 45–63, http://dx.doi.org/10.1109/9780470544785.ch2.

[28] P. Christiano, J. Leike, T.B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, 2023, arXiv:1706.03741.

[29] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C.D. Manning, C. R'e, D. Acosta-Navas, D.A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L.J. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N.S. Kim, N. Guha, N.S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S.M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T.F. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, Ann. New York Acad. Sci. 1525 (2023) 140–146, http://dx.doi.org/10.1111/nyas.15007.

[30] P.P. Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet Things Cyber-Phys. Syst. 3 (2023) 121–154, http://dx.doi.org/10.1016/j.iotcps.2023.04.003.

[31] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models, 2023, arXiv:2302.13971.

[32] A.S. Luccioni, J.D. Viviano, What's in the box? A preliminary analysis of undesirable content in the common crawl corpus, 2021, arXiv:2105.02732.

[33] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The pile: An 800GB dataset of diverse text for language modeling, 2020, arXiv:2101.00027.

[34] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, 2016, arXiv:1606.05250.

[35] A. Talmor, J. Herzig, N. Lourie, J. Berant, Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019, arXiv:1811.00937.

[36] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A human generated machine reading comprehension dataset, 2018, arXiv:1611.09268.

[37] S. Li, X. Yuan, A review of the current intelligent personal agents, 2018, pp. 253–257, http://dx.doi.org/10.1007/978-3-319-92270-6_35.

[38] G. López, L. Quesada, L.A. Guerrero, Alexa vs. Siri vs. Cortana vs. Google assistant: A comparison of speech-based natural user interfaces, in: I.L. Nunes (Ed.), Advances in Human Factors and Systems Interaction, Springer International Publishing, Cham, 2018, pp. 241–250.

[39] P. Smutný, P. Schreiberova, Chatbots for learning: A review of educational chatbots for the facebook messenger, Comput. Educ. 151 (2020) 103862, http://dx.doi.org/10.1016/j.compedu.2020.103862.

[40] F. Rakotomalala, H.N. Randriatsarafara, A.R. Hajalalaina, N.M.-V. Ravonimanantsoa, Voice user interface: Literature review, challenges and future directions, Syst. Theory, Control Comput. J. (2021) http://dx.doi.org/10.52846/stccj.2021.1.2.26.

[41] GPT3, 2021, https://openai.com/blog/gpt-3-apps. (Accessed 13 July 2023).

[42] GPT2, 2019, https://openai.com/research/gpt-2-1-5b-release. (Accessed 13 July 2023).

[43] H.R. Kirk, B. Vidgen, P. Röttger, S.A. Hale, Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback, 2023, arXiv:2303.05453.

[44] A. Holzinger, M. Kargl, B. Kipperer, P. Regitnig, M. Plass, H. Müller, Personas for artificial intelligence (AI) an open source toolbox, IEEE Access 10 (2022) 23732–23747, http://dx.doi.org/10.1109/ACCESS.2022.3154776.

[45] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A.M. Dai, A. Hauth, et al., Gemini: A family of highly capable multimodal models, 2023, arXiv preprint arXiv:2312.11805.

[46] Introducing the next generation of claude — anthropic.com, 2024, https://www.anthropic.com/news/claude-3-family. (Accessed 30 March 2024).

[47] Introducing MPT-7B: A new standard for open-source, commercially usable LLMs | databricks — databricks.com, 2024, https://www.databricks.com/blog/mpt-7b. (Accessed 30 March 2024).

[48] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, G. Penedo, The falcon series of open language models, 2023, arXiv:2311.16867.

[49] Language — Stability AI — stability.ai, 2024, https://stability.ai/stable-lm. (Accessed 30 March 2024).

[50] Long sequence modeling with XGen: A 7B LLM trained on 8K input sequence length — blog.salesforceairesearch.com, 2023, https://blog.salesforceairesearch.com/xgen/. (Accessed 30 March 2024).

[51] GPT4 blog, 2023, https://openai.com/research/gpt-4. (Accessed 13 July 2023).

[52] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT-related research and perspective towards the future of large language models, Meta-Radiol. 1 (2) (2023) 100017, http://dx.doi.org/10.1016/j.metrad.2023.100017.

[53] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P.S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, 2023, arXiv:2307.03109.

[54] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, Q. Wang, Software testing with large language models: Survey, landscape, and vision, 2024, arXiv:2307.07221.

[55] M. Chen, J. Tworek, H. Jun, Q. Yuan, H.P.d. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F.P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W.H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A.N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, W. Zaremba, Evaluating large language models trained on code, 2021, arXiv:2107.03374.

[56] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2021, arXiv:2009.03300.

[57] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, Training verifiers to solve math word problems, 2021, arXiv:2110.14168.

[58] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the MATH dataset, 2021, arXiv:2103.03874.

[59] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, HellaSwag: Can a machine really finish your sentence?, 2019, arXiv:1905.07830.

[60] Large language model LeaderBoard, 2024, https://www.vellum.ai/llm-leaderboard. (Accessed 30 March 2024).

[61] M. Arman, U. Lamiya, ChatGPT, A product of AI, and its influences in the business world, 2023, http://dx.doi.org/10.54045/talaa.v3i1.725.

[62] A. Haleem, M. Javaid, R.P. Singh, An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges, BenchCouncil Trans. Benchmarks, Stand. Eval. 2 (4) (2022) 100089, http://dx.doi.org/10.1016/j.tbench.2023.100089.

[63] G. Sebastian, Privacy and data protection in ChatGPT and other AI chatbots: Strategies for securing user information, SSRN Electron. J. (2023) http://dx.doi.org/10.2139/ssrn.4454761.

[64] H. Hassani, E.S. Silva, The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field, Big Data Cogn. Comput. 7 (2) (2023) http://dx.doi.org/10.3390/bdcc7020062.

[65] I.O. Gallegos, R.A. Rossi, J. Barrow, M.M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N.K. Ahmed, Bias and fairness in large language models: A survey, 2024, arXiv:2309.00770.

[66] Y. Li, M. Du, R. Song, X. Wang, Y. Wang, A survey on fairness in large language models, 2024, arXiv:2308.10149.

[67] S. Husse, A. Spitz, Mind your bias: A critical review of bias detection methods for contextual language models, 2022, arXiv:2211.08461.

[68] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W.Y. Wang, Mitigating gender bias in natural language processing: Literature review, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1630–1640, http://dx.doi.org/10.18653/v1/P19-1159, URL https://aclanthology.org/P19-1159.

[69] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L.A. Hendricks, W. Isaac, S. Legassick, G. Irving, I. Gabriel, Ethical and social risks of harm from language models, 2021, arXiv:2112.04359.

[70] D. Yuan, Language bias in visual question answering: A survey and taxonomy, 2021, arXiv:2111.08531.

[71] V. Thakur, Unveiling gender bias in terms of profession across LLMs: Analyzing and addressing sociological implications, 2023, arXiv:2307.09162.

[72] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Diff. 103 (2023) 102274, http://dx.doi.org/10.1016/j.lindif.2023.102274.

[73] A. Tack, C. Piech, The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues, 2022, arXiv:2205.07540.

[74] R. Rodriguez-Torrealba, E. Garcia-Lopez, A. Garcia-Cabot, End-to-end generation of multiple-choice questions using text-to-text transfer transformer models, Expert Syst. Appl. 208 (2022) 118258, http://dx.doi.org/10.1016/j.eswa.2022.118258.

[75] M. Sulaiman, K. Roy, Fair classification via transformer neural networks: Case study of an educational domain, 2022, arXiv:2206.01410.

[76] C. Potier Watkins, O. Dehaene, S. Dehaene, Automatic construction of a phonics curriculum for reading education using the transformer neural network, in: S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, R. Luckin (Eds.), Artificial Intelligence in Education, Springer International Publishing, Cham, 2019, pp. 226–231.

[77] A. Gilson, C.W. Safranek, T. Huang, V. Socrates, L. Chi, R.A. Taylor, D. Chartash, How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment, JMIR Med. Educ. 9 (2023) e45312, http://dx.doi.org/10.2196/45312.

[78] Z.A. Bukhsh, A. Saeed, R.M. Dijkman, ProcessTransformer: Predictive business process monitoring with transformer network, 2021, arXiv:2104.00721.

[79] T. Douzon, S. Duffner, C. Garcia, J. Espinas, Improving information extraction on business documents with specific pre-training tasks, in: S. Uchida, E. Barney, V. Eglin (Eds.), Document Analysis Systems, Springer International Publishing, Cham, 2022, pp. 111–125.

[80] M. Heidari, S. Rafatirad, Semantic convolutional neural network model for safe business investment by using BERT, in: 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS, 2020, pp. 1–6, http://dx.doi.org/10.1109/SNAMS52053.2020.9336575.

[81] A. Agarwal, S. Maiya, S. Aggarwal, Evaluating empathetic chatbots in customer service settings, 2021, arXiv:2101.01334.

[82] S. Humeau, K. Shuster, M.-A. Lachaux, J. Weston, Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring, 2020, arXiv:1905.01969.

[83] S. Yadav, D. Gupta, A.B. Abacha, D. Demner-Fushman, Question-aware transformer models for consumer health question summarization, J. Biomed. Inform. 128 (2022) 104040, http://dx.doi.org/10.1016/j.jbi.2022.104040.

[84] Y. Li, S. Rao, J.R.A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, G. Salimi-Khorshidi, BEHRT: Transformer for electronic health records, Sci. Rep. 10 (1) (2020) 7155, http://dx.doi.org/10.1038/s41598-020-62922-y.

[85] M. Polignano, M. de Gemmis, G. Semeraro, et al., Comparing transformer-based NER approaches for analysing textual medical diagnoses, 2021.

[86] W. Lin, P. Babyn, W. Zhang, et al., Context-based ontology modelling for database: Enabling ChatGPT for semantic database management, 2023, arXiv preprint arXiv:2303.07351.

[87] Y. Song, J. Mahmud, N. De Silva, Y. Zhou, O. Chaparro, K. Moran, A. Marcus, D. Poshyvanyk, BURT: A chatbot for interactive bug reporting, 2023, arXiv preprint arXiv:2302.06050.

[88] Global statistics on chatbots, 2022, https://www.statista.com/statistics/1007392/worldwide-chatbot-market-size/. (Accessed 13 July 2023).

[89] S. Zainol, M.F. Shamsudin, S. Hassan, N.A. Mohd Noor, Understanding customer satisfaction of chatbots service and system quality in banking services, J. Inf. Technol. Manag. 15 (Special Issue) (2023) 142–152, http://dx.doi.org/10.22059/jitm.2022.89417.

[90] A. Borji, A categorical archive of ChatGPT failures, 2023, arXiv:2302.03494.

[91] J. Zhou, H. Müller, A. Holzinger, F. Chen, Ethical ChatGPT: Concerns, challenges, and commandments, 2023, arXiv:2305.10646.

[92] A. Rao, S. Vashistha, A. Naik, S. Aditya, M. Choudhury, Tricking LLMs into disobedience: Understanding, analyzing, and preventing jailbreaks, 2023, arXiv:2305.14965.

[93] M. Zong, B. Krishnamachari, A survey on GPT-3, 2022.

[94] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, 2021, arXiv:2005.11401.

[95] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, 2021.

[96] G.K.-M. Liu, Perspectives on the social impacts of reinforcement learning with human feedback, 2023, arXiv:2303.02891.

[97] D. Ganguli, A. Askell, N. Schiefer, T.I. Liao, K. Lukošiūtė, A. Chen, A. Goldie, A. Mirhoseini, C. Olsson, D. Hernandez, D. Drain, D. Li, E. Tran-Johnson, E. Perez, J. Kernion, J. Kerr, J. Mueller, J. Landau, K. Ndousse, K. Nguyen, L. Lovitt, M. Sellitto, N. Elhage, N. Mercado, N. DasSarma, O. Rausch, R. Lasenby, R. Larson, S. Ringer, S. Kundu, S. Kadavath, S. Johnston, S. Kravec, S.E. Showk, T. Lanham, T. Telleen-Lawton, T. Henighan, T. Hume, Y. Bai, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, C. Olah, J. Clark, S.R. Bowman, J. Kaplan, The capacity for moral self-correction in large language models, 2023, arXiv:2302.07459.

[98] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, Toxicity in ChatGPT: Analyzing persona-assigned language models, 2023, arXiv:2304.05335.

[99] E.K. Tokpo, T. Calders, Text style transfer for bias mitigation using masked language modeling, 2022, arXiv:2201.08643.

[100] A. Wei, N. Haghtalab, J. Steinhardt, Jailbroken: How does LLM safety training fail?, 2023, arXiv:2307.02483.

[101] M. Gupta, C. Akiri, K. Aryal, E. Parker, L. Praharaj, From ChatGPT to threatgpt: Impact of generative AI in cybersecurity and privacy, 2023, arXiv:2307.00691.

[102] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, Y. Liu, Jailbreaking ChatGPT via prompt engineering: An empirical study, 2023, arXiv:2305.13860.

[103] N. Meade, S. Gella, D. Hazarika, P. Gupta, D. Jin, S. Reddy, Y. Liu, D. Hakkani-Tür, Using in-context learning to improve dialogue safety, 2023, arXiv:2302.00871.

[104] U.P. Liyanage, N.D. Ranaweera, Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts, J. Comput. Soc. Dyn. 8 (11) (2023) 15–25.

[105] K.-J. Tokayev, Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns, Int. J. Soc. Anal. 8 (9) (2023) 17–33.

[106] W. Kidder, J. D'Cruz, K.R. Varshney, Empathy and the right to be an exception: What LLMs can and cannot do, 2024, arXiv:2401.14523.

[107] A. Rao, A. Khandelwal, K. Tanmay, U. Agarwal, M. Choudhury, Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs, 2023, arXiv:2310.07251.

[108] H. Müller, M.T. Mayrhofer, E.-B. Van Veen, A. Holzinger, The ten commandments of ethical medical AI, Computer 54 (7) (2021) 119–123.