IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 6 October 2023, accepted 22 November 2023, date of publication 5 December 2023,
date of current version 22 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3339553

SURVEY

Unleashing the Potential of Conversational AI: Amplifying Chat-GPT's Capabilities and Tackling Technical Hurdles

VIKAS HASSIJA[1], ARJAB CHAKRABARTI[2], ANUSHKA SINGH[2], VINAY CHAMOLA[3,4], (Senior Member, IEEE), AND BIPLAB SIKDAR[5], (Senior Member, IEEE)

[1]School of Computing, National University of Singapore, Singapore 117417
[2]School of Computer Science, KIIT, Bhubaneshwar 751024, India
[3]Department of Electrical and Electronics Engineering, BITS Pilani, Pilani Campus, Pilani 333031, India
[4]APPCAIR, BITS Pilani, Pilani Campus, Pilani 333031, India
[5]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117417

Corresponding author: Vinay Chamola (vinay.chamola@pilani.bits-pilani.ac.in)

ABSTRACT Conversational AI has seen a growing interest among government, researchers, and industrialists. This comprehensive survey paper provides an in-depth analysis of large language models, specifically focusing on ChatGPT. This paper discusses the architecture, training process, and challenges associated with large language models, including bias, interpretability, and ethics. It explores various applications of ChatGPT and examines future research trends, such as improving model generalization, addressing data scarcity, and integrating multimodal capabilities. This survey also serves as a roadmap for researchers, practitioners, and policymakers, offering valuable insights into the current state and future potential of large language models and ChatGPT.

INDEX TERMS Large language models, ChatGPT, natural language processing, deep learning, neural networks, transformer models, pre-training and fine-tuning, language generation, text completion, model interpretability, bias in language models, ethics in AI, data scarcity, multimodal models, generalization, conversational AI, language understanding, text classification, sentiment analysis, dialogue systems.

I. INTRODUCTION

In the recent years, Large Language Models (LLM) have revolutionized Natural Language Processing (NLP) by demonstrating exceptional capabilities in understanding and generating text that closely resembles human language [1]. Equipped with deep learning techniques and trained on extensive datasets, these LLM models have led to significant advancements in various NLP applications such as machine translation [2], question answering [3], text summarization [4], and conversational agents [5]. ChatGPT, developed by OpenAI, is widely acknowledged as one of the foremost exemplars of large language models. It has garnered considerable attention from both the general public and researchers. The system's impressive capability to generate coherent and contextually appropriate responses has been a pivotal factor in its widespread acclaim [6], [7].

Continuing the progress in bridging the gap between machines and human language understanding, large language models have emerged as a powerful solution to address the limitations of traditional NLP techniques. These models employ deep learning algorithms and vast amounts of text data to learn patterns and relationships within language, enabling them to generate high-quality, context-aware text [8].

One of the primary motivations behind the development of large language models is their ability to learn from extensive corpora [9]. Unlike rule-based approaches that depend on predefined patterns and crafted features, large language models benefit from training on copious amounts

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

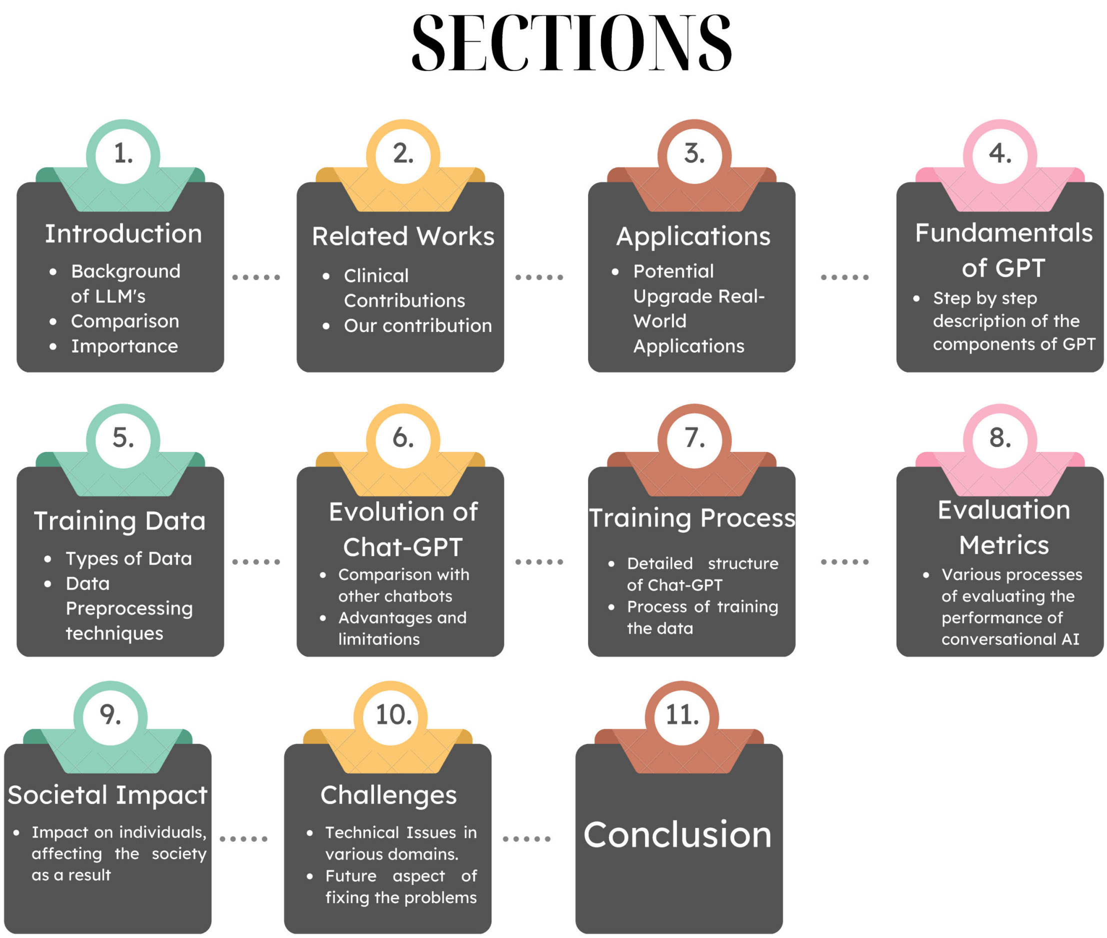


FIGURE 1. Sections of the paper.

of text data, including books, articles, websites, and other sources [10]. Therefore, by adopting a data-driven approach, these models can capture the intricacies and subtleties of human language [11]. As a result, they produce more accurate and contextually relevant responses.

The potency of deep neural networks utilized in these models empowers them to comprehend and generate text at a more nuanced level [12]. The models acquire representations of words and phrases in a continuous vector space, allowing them to capture semantic relationships and contextual dependencies. Therefore, by learning from the diverse textual information, large language models develop a comprehensive understanding of language, enabling them to generate coherent and natural-sounding responses [13].

Moreover, the ability of large language models to generate context-aware text is a crucial aspect of their success [14]. These models learn from the provided context in the input prompt or query and produce responses that consider the preceding information. This contextual understanding enables the models to provide more accurate and appropriate answers, leading to more meaningful and engaging user interactions.

The availability of large language models like ChatGPT has opened up new possibilities for various NLP applications. These models can be integrated into chatbots [15], virtual assistants, and customer service systems [16], enabling more effective and human-like interactions with users. These models can assist in language translation, summarization, and information retrieval tasks, offering faster and more precise results. The development of large language models has significantly advanced the field of NLP, offering a promising pathway towards achieving improved human-machine communication [17].

The need for this survey paper arises from the rapid advancements and growing interest in chatbot technology [18], explicitly focusing on the evolution and impact of the GPT architecture [19]. As chatbots play a significant role in various domains, understanding the development, strengths, and limitations of different chatbot models becomes crucial [20]. This survey paper aims to provide a comprehensive overview and analysis of the evolution of chatbot models, with a specific emphasis on integrating the GPT architecture [21]. By examining the early developments in conversational AI, comparing different chatbot models, and highlighting the advancements brought by GPT-based

TABLE 1. Comparison of various large language models.

Model	Organization	Pretraining Architecture	Parameters	Released	Use Cases
BERT	Google	Encoder	340 Million	2018	Language understanding, question-answering
GPT	Open-AI	Decoder	117 Million	2018	Language generation, text completion
GPT-2	Open-AI	Transformer	1.5 Billion	2019	Language generation, text completion
GPT-3	Open-AI	Transformer	175 Billion	2020	Language generation, translation, question-answering, text completion
T5	Google	Transformer	11 Billion	2020	Language generation, translation, question-answering, text completion
DALL-E	Open-AI	Transformer	12 Billion	2021	Image generation from textual input
CLIP	Open-AI	Transformer	33 Million	2021	Image and text understanding
Codex	Open-AI	Transformer	12 Billion	2021	Code generation from natural language input
LaMDA	Google	Transformer	137 Billion	2021	Conversational AI
GShard	Google	Transformer	600 Billion	2021	Language generation, translation, question-answering, text completion
Chinchilla	Microsoft	Transformer	17 Billion	2021	Language generation, text completion
FLAN	Facebook	Transformer	2.6 Billion	2021	Language generation, translation, question-answering, text completion
Alpaca	Stanford	Transformer	7 Billion	2021	Language generation, text completion
GPT-Neo	EleutherAI	Transformer	2.7 - 2.8 Billion	2021	Language generation, text completion
Chat-GPT	Open-AI	Fine-Tuning	175 Billion	2022	Conversational AI
Flamingo	Google	Transformer	80 Billion	2022	Conversational AI
BLOOM	BigScience	Transformer	1.6 Trillion	2022	Multilingual language understanding, open science, open access
GPT-4	Open-AI	Transformer	1 Trillion	2023	Content creation, Language translation, Customer service and Chatbot, Healthcare contract management
Chinchilla 2.0	Microsoft	Transformer	34 Billion	2023	Conversational AI
LLaMA	Meta AI	Transformer	7 B - 65 Billion	2023	Language understanding, natural language processing
BLOOM 2.0	BigScience	Transformer	3.2 Trillion	2023	Conversational AI

approaches, the paper aims to shed light on the progress made in dialogue modelling and response generation.

Our work stands out from other reviews or surveys on ChatGPT due to the following unique aspects:

– **Comprehensive coverage:** Our survey paper provides a comprehensive overview of the evolution of large language models, explicitly focusing on integrating the GPT architecture [22]. We cover early developments in conversational AI, compare ChatGPT with earlier models, and explore advancements in dialogue modelling and response generation.

– **In-depth analysis:** We offer in-depth analysis and insights into the strengths and limitations of LLM's, focusing specifically on ChatGPT [23]. We examine

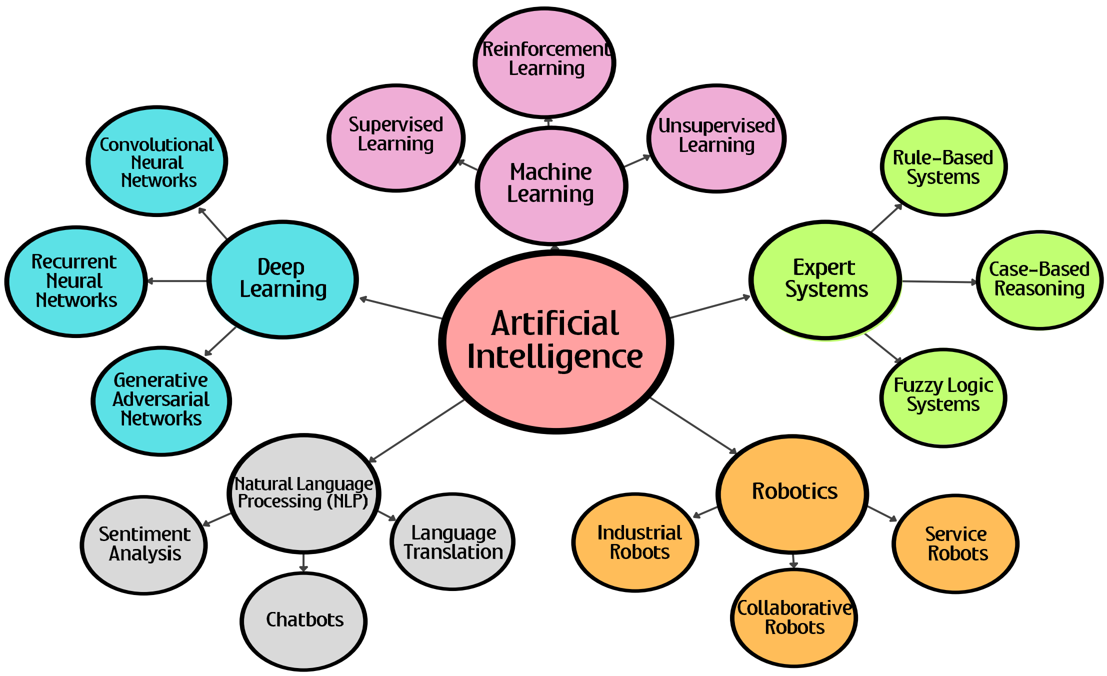


FIGURE 2. Types of artificial intelligence.

its contextual understanding, coherence, long-range dependency modelling, and handling of ambiguous queries or user intents [24].

– **Critical evaluation:** Our survey paper critically evaluates Chat-GPT, an LLM created by Open AI, by addressing issues related to response quality control [20], bias mitigation, and scalability. This evaluation provides a balanced perspective on the practical considerations of deploying ChatGPT.
– **Focus on relevance and application:** While covering technical aspects, we also emphasize the relevance and application of LLM's in real-world scenarios [25] and the implications of ChatGPT's advancements in dialogue modeling [26].

The intended audience for this survey paper includes researchers, practitioners, and enthusiasts in conversational AI and natural language processing [27]. It provides a comprehensive overview and analysis of the evolution of chatbot models, focusing on integrating the GPT architecture [22]. Researchers will gain valuable insights into dialogue modelling and response generation advancements, while practitioners will benefit from practical considerations and implementation guidance. Enthusiasts interested in conversational AI will find accessible information about the development, strengths, and limitations of chatbot models, particularly for ChatGPT.

II. RELATED WORKS

The field of data visualization and Natural Language Interfaces (NLIs) has seen significant research and development in recent years. Researchers have explored various techniques to generate visualizations directly from natural language text, aiming to bridge the gap between textual queries and visual representations [31]. Previous studies have focused on refining NLIs, addressing challenges posed by the ambiguity of natural language and poorly written user queries. These studies have primarily relied on traditional approaches that utilize hand-crafted grammar rules and tailored models. However, these methods often suffer from limitations in accuracy and complexity. In contrast, Hiroshi Honda et al. [28] proposed the fusion of symbolic processing with deep learning. By utilizing techniques like Neural Machine Translation (NMT) and Word2Vec, they successfully emulate Prolog systems, which are traditionally used for symbolic processing. Their novel approach stands out for its ability to manage unknown data, presenting a promising avenue for developing question-answering systems built atop Prolog knowledge bases.

Luis Martín Sánchez-Adame et al. [18] emphasized the undervalued importance of usability evaluations. Recognizing a gap in the existing heuristic evaluations tailored for chatbots, they introduced five specific usability heuristics. These were born from both hands-on experience in chatbot development and a comprehensive review of existing literature.

Paula Maddigan et al. [29] proposed a novel system called Chat2VIS that leverages the capabilities of large language models (LLMs) such as ChatGPT and GPT-3 to convert free-form natural language into code for generating appropriate visualizations. By employing effective, prompt engineering, the study demonstrates the potential for LLMs to provide more straightforward, more accurate end-to-end solutions for rendering visualizations. Additionally, this

TABLE 2. Related works.

Author(s) and Reference	Summary	Advantages	Disadvantages
Hiroshi Honda *et al.* [28] [2019]	Proposed methods for symbolic processing learning with deep learning and building question answering systems.	- Superior handling of unknown data. - Construction of knowledge bases.	- Complexity of implementing symbolic processing with deep learning. - Limited generalization to complex reasoning tasks.
Luis Martín Sánchez-Adame *et al.* [18] [2021]	Focused on usability evaluations of chatbots and proposed specific usability heuristics. Conducted a case study to validate the proposed heuristics in an education-oriented chatbot.	- Improved usability of chatbots. - Specific usability heuristics for evaluation. - Validation through case study.	- Usability heuristics may not cover all scenarios. - Limited generalizability to different chatbot domains.
Paula Maddigan *et al.* [29] [2023]	Proposed Chat2VIS system leveraging large language models (LLMs) for converting natural language into code for visualizations.	- Cost reductions - Improved accuracy - Adaptability to different datasets	- Potential loss of control over the generated code - Dependency on the availability and quality of large language models
Volker Bilgram *et al.* [21] [2023]	Explored the potential of LLMs, specifically GPT, in revolutionizing innovation management. Provided insights into human-AI collaboration.	- Revolutionize innovation management - Use cases in user journey mapping, idea generation, and digital prototyping - Insights into human-AI collaboration	- Ethical concerns and biases in AI-generated ideas - Overreliance on AI for decision making
Tianyu Wu *et al.* [30] [2023]	Provided an overview of ChatGPT, discussing its techniques, advantages, and potential future development. Emphasized the importance of responsible use and highlighted open problems for future research.	- Powerful language generation capabilities - Wide range of applications - Potential for future development	- Limited control over generated output - Potential for biases and misinformation
Our work [2023]	Described our survey paper on conversational AI and ChatGPT, providing an in-depth analysis of the current state-of-the-art techniques, applications, challenges and future directions.	- Comprehensive overview of conversational AI and ChatGPT - In-depth analysis of techniques and applications - Identification of key challenges	- Limited focus on specific implementation details - Potential bias in the selection of surveyed works

approach offers significant cost reductions in NLI system development while maintaining superior visual inference abilities. This study also addresses concerns regarding data security and privacy by showcasing the construction of LLM prompts adaptable to different datasets.

Volker Bilgram et al. [21] introduces an article that provides valuable insights from six months of experimentation, highlighting generative AI's role in areas such as user journey mapping and early prototyping. Their findings not only underscore the efficiency and cost-effectiveness of AI-assisted innovation but also delve into the dynamics of human-AI collaboration in the innovation process.

Tianyu Wu et al. [30] introduces a paper that provides an overview of ChatGPT, an artificial intelligence-generated content (AIGC) model developed by OpenAI. The paper dissected the model's history, current status, and prospective developments. Their analysis touches upon the core techniques underpinning ChatGPT, weighing its advantages against potential challenges. While acknowledging the myriad opportunities it presents, they also stress the responsible usage of such models to mitigate academic and safety concerns.

In conclusion, the studies reviewed in this section have significantly contributed to the field of chatbot usability evaluations. They have highlighted the growing importance of considering usability in chatbot development and the potential challenges that arise from rapid development processes. The papers discussed above represents an important step forward in addressing these challenges. These findings expand the horizon of usability evaluations in the context of chatbots, offering valuable insights for developers and researchers alike. Moving forward, further research and refinement of the suggested guidelines are necessary to improve their applicability and effectiveness in assessing chatbot usability. By doing so, we can ensure that chatbots are designed to provide optimal user experiences, leading to improved adoption and satisfaction among users.

III. APPLICATIONS OF LARGE LANGUAGE MODEL

Powerful Large Language Models(LLMs) like ChatGPT (based on the GPT-4 architecture), have opened up new possibilities for improving customer service operations and enhancing human-computer interactions [32]. This section explores the diverse applications of a famous large language model, ChatGPT, in customer service and support, response time and accuracy, language translation and understanding, text generation and summarization, personal assistants, social chatbots, and entertainment. Aditionally, by examining existing research and industry insights, we provide a comprehensive analysis of LLM's potential in transforming the above mentioned fields and beyond.

A. CHATBOT DEPLOYMENT IN CUSTOMER SERVICE OPERATIONS

One significant application of LLM's lies in its deployment as a chatbot in customer service operations. Chatbots powered by LLM's like ChatGPT can significantly enhance customer interactions and streamline support processes. By integrating them into customer-facing chatbot platforms, organizations can benefit from its language generation capabilities and ability to provide prompt and accurate responses to customer queries [33].

It enables organizations to handle a higher volume of customer inquiries simultaneously, reducing the wait time for customers and improving overall response time. With ChatGPT's ability to understand and generate human-like text, it can effectively interact with customers in a conversational manner, offering a more natural experience than traditional automated systems.

It's language generation capabilities also contribute to improving the accuracy of customer interactions. It can generate responses that are contextually relevant and accurate, based on the information provided by customers. Furthermore, integrating ChatGPT into customer service chatbots enables organizations to enhance their customer support across various communication channels, such as email, social media, and product review websites. ChatGPT can analyze and respond to customer complaints, reviews, and inquiries on different platforms, providing consistent and efficient support.

B. LANGUAGE TRANSLATION AND UNDERSTANDING

LLM's offers a valuable application in the realm of language translation and understanding. Their language generation capabilities and comprehensive training data make it a useful tool for overcoming language barriers and enhancing communication between individuals who speak different languages. The integration of ChatGPT into translation tools or chat interfaces can provide multilingual conversational agents that assist customers in their preferred language.

Language translation powered by LLM's not only improves the accessibility of communication but also enhances the overall experience by eliminating language-related misunderstandings and facilitating clear dialogue. Customers can express their queries and concerns in their native language, while support agents receive translated versions of those messages, enabling them to respond accurately. Furthermore, LLM's language understanding capabilities help address the challenges posed by idiomatic expressions and language nuances. It can interpret and provide appropriate responses, taking into account the cultural context and specific linguistic characteristics. This enables more accurate and contextually relevant interactions, contributing to enhanced satisfaction [34].

It is important to note that while LLM's can facilitate language translation and understanding, its accuracy and proficiency in handling complex or specialized terminology may have limitations. Organizations should evaluate and fine-tune the translation models, considering specific industry jargon or domain-specific language requirements. Furthermore, leveraging ChatGPT's language translation and understanding capabilities, organizations can provide seamless communication to individuals across different languages, expanding their reach and improving overall satisfaction.

C. TEXT GENERATION AND SUMMARIZATION

The text generation and summarization capabilities of LLM's offer valuable applications in various domains, allowing organizations to generate coherent and contextually relevant text while summarizing lengthy documents or conversations. The ability of ChatGPT to produce human-like text can be utilized to compose responses, emails, and informative support documentation. This empowers users, such as customer service agents, to provide personalized and informative messages in a clear and concise manner. Leveraging their text generation capabilities streamlines operations and ensures consistent and high-quality responses.

Additionally, ChatGPT's summarization abilities prove useful in condensing lengthy documents, articles, or conversations into concise summaries. This is particularly beneficial when there is a need to review extensive information quickly [35]. The application of text generation and summarization extends beyond customer service, finding utility in automated email responses and chatbot interactions. Similarly, when integrated into chatbot platforms, they can

TABLE 3. Applications of LLMs in various fields.

Field	Applications of LLMs	Benefits	Challenges	Examples
Language Translation	Neural machine translation; Contextual embeddings; Transformer models	High accuracy; Real-time translations	Idiomatic expressions; Zero-shot translation	BERT; Transformer NMT
Customer Service Chatbot	Intent recognition; Slot filling; Context-aware response	Reduced manual operations; Improved resolution	Continual training; Context persistence	GPT chatbots
Text Generation	Abstractive summarization; Controlled generation; Fine-tuning	Consistent summaries; Scalability	Incorrect info; Domain jargon	GPT summarization
Personal Assistants	Voice-command recognition; Task scheduling; Contextual reminders	IoT integration; Proactive management	Multimodal data; Overlapping commands	Siri; Google Assistant
Social Chatbots	Dynamic storylines; Procedural content; Personality interactions	User immersion; Personalized experiences	Appropriate content; Real-time performance	LLM NPCs
Innovation Tools	Prototyping; Code generation; Data extraction	Rapid prototyping; Language to code	Code security; Task precision	GPT code generation

summarize queries or provide concise responses, improving the efficiency and user experience of chatbot interactions.

D. PERSONAL ASSISTANTS AND VIRTUAL COMPANIONS

One of the compelling applications of LLM's is their potential to serve as a personal assistant or virtual companion, offering assistance, organization, and even emotional support to individuals [36]. For example, Integrating ChatGPT into personal productivity tools or virtual assistant platforms can provide users and organizations with an intelligent and interactive companion that assists with various tasks and activities [37].

LLM's can contribute to personal productivity and organization by acting as a virtual assistant that helps manage schedules, set reminders, and provide timely notifications. It can assist users in organizing their tasks, appointments, and deadlines, increasing efficiency and reducing cognitive load. Through natural language interactions, LLM's can understand user preferences and generate personalized recommendations, assisting individuals in making informed decisions [38].

Furthermore, their potential as a virtual companion extends beyond practical tasks. It can offer emotional support and companionship, providing individuals with someone to talk to and share their thoughts and feelings [39].

It is crucial to consider the ethical implications and potential limitations associated with virtual companions powered by LLM's like ChatGPT. While they can provide valuable support, there is a need to ensure transparency about their artificial nature and maintain clear boundaries regarding the scope of their capabilities. Care should also be taken to address potential biases or unintended reinforcement of negative behaviors that may arise from the interaction with virtual companions.

E. SOCIAL CHATBOTS AND ENTERTAINMENT

LLM's language generation capabilities make it an ideal tool for developing social chatbots that can provide interactive and engaging conversational experiences for users. They can be deployed in various social platforms and applications, offering entertainment, companionship, and personalized interactions [40]. Social chatbots powered by ChatGPT can simulate human-like conversations, allowing users to engage in dialogues, ask questions, and receive responses that create an immersive and entertaining experience. They can be designed to embody specific personalities or characters, enabling users to interact with virtual entities in a social and interactive manner [41]. These chatbots can offer a range of entertainment experiences, such as storytelling, interactive games, trivia challenges, or role-playing scenarios. Additionally, social chatbots can provide companionship and emotional support [42], acting as virtual friends or confidants for users. Through conversational interactions, these chatbots can offer empathetic responses, engage in casual conversations, and even provide advice or encouragement, catering to the emotional needs of users seeking companionship or support. They have the potential to revolutionize the way users engage with technology. It creates opportunities for interactive storytelling experiences, immersive gaming scenarios, and social interactions that enhance user satisfaction and enjoyment [43].

However, it is crucial to establish clear boundaries and guidelines for social chatbots to ensure user safety and well-being. Organizations should implement appropriate safeguards and monitoring mechanisms to prevent misuse or exploitation of chatbot interactions. Regular evaluation and testing of the chatbot's responses and behaviors are necessary to ensure they align with ethical standards and do not promote harmful or inappropriate content.

F. VERSATILE TOOL FOR INNOVATION

LLM's offers a range of capabilities that can benefit researchers and developers in various domains. Their cability to retrieve information, generate ideas, assist in prototyping and design, handle language-related tasks, simulate scenarios, and facilitate collaboration and communication make it a valuable tool in the Research and Development (R&D) process. Research and development (R&D) is a dynamic process that requires continuous exploration, innovation, and problem-solving. In recent years, AI language models have emerged as powerful tools for supporting R&D efforts [44]. This section focuses on the utilization of Chat-GPT, a very famous large language model, as an aid for researchers and developers in their R&D work.

- **Information Retrieval:** ChatGPT possesses an extensive pre-trained knowledge base that encompasses a wide array of topics. Leveraging this knowledge, researchers can utilize ChatGPT to retrieve relevant information for literature reviews, background research, and acquiring specific facts and figures pertinent to their R&D projects [45].
- **Idea Generation:** The ability of ChatGPT to generate novel ideas and provide inspiration proves invaluable during the ideation phase of R&D. Researchers can engage in brainstorming sessions with ChatGPT, leveraging its alternative perspectives, suggestions for possible approaches, and exploration of diverse angles or possibilities [21].
- **Prototyping and Design:** ChatGPT can facilitate the prototyping and design processes by offering feedback, suggesting improvements, and providing creative solutions. Researchers and developers can refine their concepts, explore different design options, and address potential challenges with the aid of ChatGPT.
- **Language-related Tasks:** The versatility of ChatGPT extends to a range of language-related tasks encountered in R&D work [46]. From writing drafts, summarizing research papers, and proofreading to generating code snippets and assisting with language translation or text generation, ChatGPT can serve as a valuable resource for researchers and developers.
- **Simulation and Modeling:** ChatGPT can simulate scenarios or models based on the input provided by researchers [47]. This capability allows for the generation of hypothetical scenarios, virtual experimentation, and exploration of potential outcomes, contributing to the iterative nature of the R&D process.
- **Collaboration and Communication:** The collaborative potential of ChatGPT can foster improved communication and engagement within research teams [48]. It acts as a virtual collaborator or assistant, providing answers to questions, explanations, and engaging in discussions, thus enhancing collaboration and promoting efficient knowledge exchange.

ChatGPT offers a wide range of functionalities that can greatly support researchers and developers in their R&D work. However, it is important to exercise caution and verify critical information obtained through these LLM's using reliable sources and domain experts.

The next section discusses the Fundamentals of Generative Pre-trained Transformer and their groundbreaking impact on natural language processing (NLP) along with their future aspect. Understanding the fundamentals of GPT is essential for grasping the underlying principles that drive their remarkable language generation capabilities.

IV. FUNDAMENTALS OF GENERATIVE PRE-TRAINED TRANSFORMER (GPT)

Generative Pre-trained Transformer (GPT) employs a deep neural network architecture comprising multiple layers of self-attention and feed-forward neural networks [49]. This intricate design allows the model to capture complex patterns and dependencies within the input data. The self-attention mechanism enables the model to focus on different parts of the input sequence during processing, enabling it to grasp the relationships between different words or tokens. Furthermore, as a large language model, GPT possesses immense capacity to store and retrieve vast amounts of information, making it adept at understanding context and producing detailed responses. The feed-forward neural networks further enhance the model's ability to learn and generalize by applying nonlinear transformations to the representations obtained from the self-attention mechanism. Together, these components enable GPT to generate coherent and contextually relevant language outputs [50]. It employs unsupervised pre-training followed by fine-tuning on specific tasks to enhance its performance and adaptability [51].

The key components of transformer are mentioned as follows:

- **Self-Attention Mechanism:** The self-attention mechanism is a key component of the transformer model and GPT. It allows the model to attend to different parts of the input sequence, capturing relationships and dependencies between words. Self-attention enables

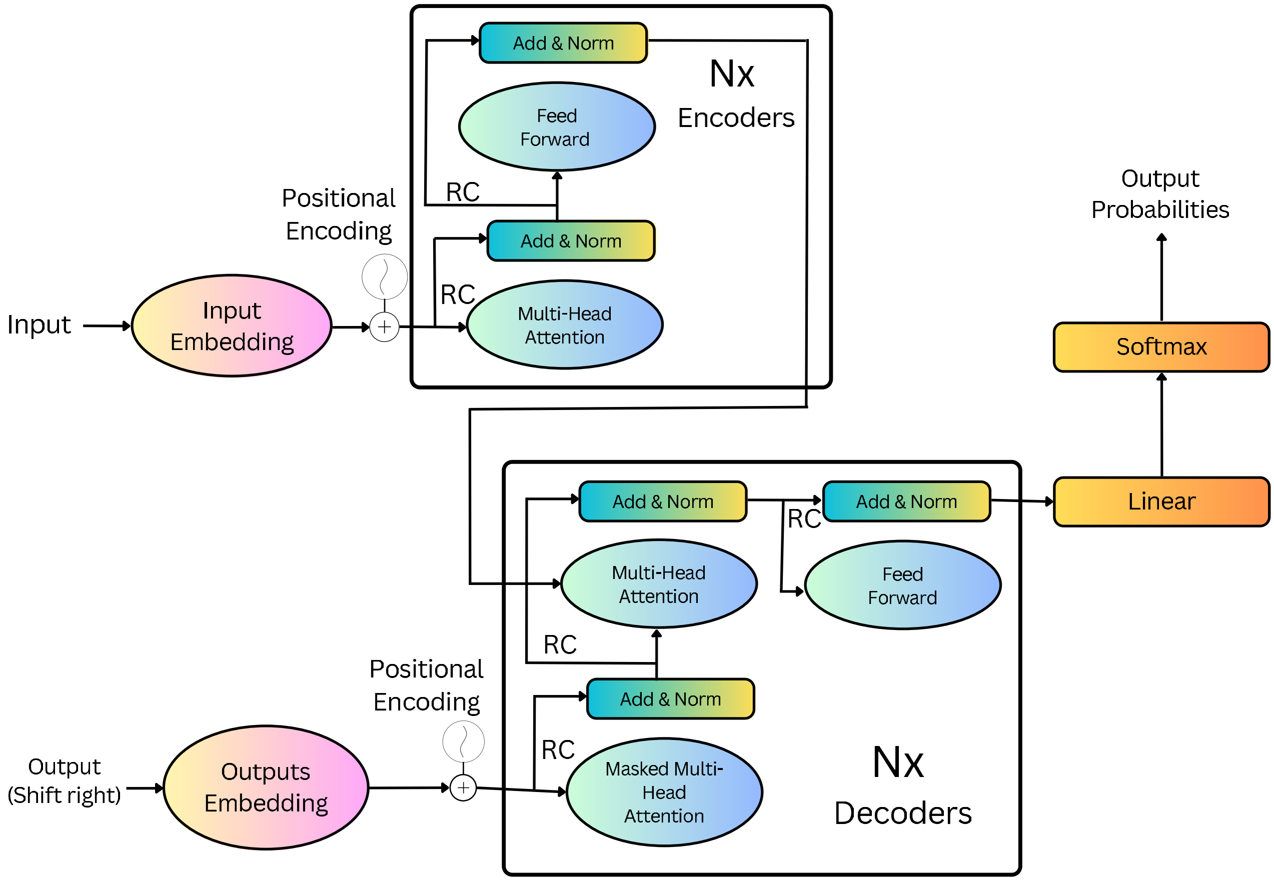


FIGURE 3. Transformer architecture.

GPT to model long-range dependencies effectively and learn contextual representations of words.

- **Positional Encoding:** To incorporate the sequential nature of language, GPT employs positional encoding. Positional encodings are added to the input embeddings to convey the order of words in a sentence. This helps the model differentiate between words with similar embeddings but different positions. Positional information is crucial for the model to understand the sequential structure of the input and generate coherent and contextually appropriate responses.
- **Feed-Forward Neural Networks:** The GPT architecture consists of multiple layers of self-attention and feed-forward neural networks. It employs feed-forward neural networks within each layer of the transformer. These networks process the outputs of the self-attention mechanism and apply non-linear transformations, enabling the model to capture complex patterns and dependencies in the data. Each layer can be seen as a transformer block, and the number of layers can vary depending on the specific GPT model. The primary goal of GPT is to generate human-like responses given an input prompt.
- **Pre-training Process of GPT:** The architecture of GPT is specifically crafted for unsupervised pre-training, which is then followed by fine-tuning for specific downstream tasks [52]. During the pre-training phase, the model is exposed to a diverse and large-scale corpus, allowing it to grasp intricate language patterns. One key technique utilized is Masked Language Modeling (MLM), where GPT tries to predict words that are intentionally hidden. In addition to MLM, GPT employs the Next Sentence Prediction (NSP) objective. This approach helps the model better understand the broader context of sentences and ensures that generated text is both meaningful and coherent.
- **Fine-Tuning and Adaptation for Specific Tasks:** After pre-training, GPT undergoes fine-tuning on task-specific datasets. This process allows the model to specialize in particular NLP tasks, such as language translation, sentiment analysis, or question answering.
- **Transfer Learning and Domain Adaptation:** GPT's pre-training enables effective transfer learning, where the knowledge learned from a large corpus can be transferred to downstream tasks [53]. This transfer learning capability allows GPT to adapt to different

domains and tasks with minimal training data, leading to improved performance and efficiency [54].
- **Techniques for Model Compression and Efficiency:** To address the computational demands of GPT, researchers have explored techniques for model compression and efficiency. Methods such as knowledge distillation [55], parameter pruning [56], and quantization [57] have been employed to reduce the model's size and computational requirements while maintaining its performance.

Generative Pre-trained Transformers (GPT) have continued to evolve and advance with subsequent models that surpass their predecessors in terms of model size, capabilities, and performance. This section provides an overview of notable variations and advancements such as GPT-2, GPT-3, GPT-3.5 and GPT-4.

1) GPT-2, introduced in 2019, represented a significant leap in performance compared to the original GPT model [61]. It featured a substantial increase in model size, reaching up to 1.5 billion parameters. GPT-2 showcased impressive language generation capabilities, generating coherent and contextually relevant text. The large languaage model gained attention for its versatility across various language tasks, including text completion, summarization, and story generation.
2) GPT-3, released in 2020, pushed the boundaries of language generation even further [29]. It achieved a remarkable model size of 175 billion parameters, making it one of the largest language models ever created. GPT-3 had exhibited exceptional language understanding and generation abilities, generating diverse and contextually rich responses. The LLM demonstrated remarkable performance across various tasks and domains, showcasing its potential for applications such as translation, question answering, and creative writing.
3) GPT-3.5 represents a further advancement based on GPT-3 [62]. It builds upon the strengths of GPT-3, with improvements in training methodologies, architecture, and efficiency. GPT-3.5 enhances language understanding and generation quality, offering refined control over the generated outputs. It addresses some limitations of earlier models and refines the overall user experience.
4) GPT-4 is the next major iteration of the GPT series [63]. GPT-4 is the latest version of the Generative Pre-trained Transformer (GPT) series of large language models. It is a large multi-modal model that can process both image and text inputs and produce text outputs [64]. GPT-4 is a significant improvement on GPT-3, outperforming other models in English and far outperforming them in other languages. One of the most significant differences between GPT-4 and its predecessors is that GPT-4 can process images alongside text, making it a multimodal model.

As GPT models have become more powerful, there has been increased attention on addressing ethical concerns and ensuring proper control over generated content. Advancements are being made to enhance filtering mechanisms, promote responsible AI use, and provide users with more control and transparency in guiding the output of GPT models. Continued research focuses on reducing biases [65], addressing potential harmful content generation, and enabling users to shape the behavior and values of the models. Beyond the aforementioned models, researchers and practitioners have explored variations and specialized versions of GPT. This includes domain-specific fine-tuning [66] to tailor GPT models for specific industries or tasks, such as healthcare, legal, or financial domains [46]. These variations aim to improve performance and cater to specific requirements of various applications [67].

Research on GPT and subsequent models is an active and rapidly evolving field. Ongoing efforts are focused on improving training efficiency, model compression techniques, addressing biases and limitations, and exploring novel approaches to enhance language understanding and generation. The future direction of GPT and subsequent models is likely to involve advancements in performance, scalability, interpretability, and control [68].

In the next section, we delve into the Training Data and Corpus of Generative Pre-trained Transformers (GPT) and their profound impact on natural language processing (NLP), as well as their future implications.

V. TRAINING DATA AND CORPUS

In the field of conversational AI and training large language models models like ChatGPT, the selection and collection of conversational data for pre-training is a critical step [69]. Various sources of conversational data can be used, including social media platforms, online forums, chat logs, and even specific datasets created for research purposes [70]. Considerations for data quality, relevance, and diversity are essential to ensure the effectiveness and generalizability of the trained models.

The size and composition of the corpus may vary depending on the specific requirements of the pre-training task and the target application. For example, some studies use a combination of public datasets, such as Wikipedia, Reddit, or Twitter, while others collect and curate their own conversational datasets. The data is often filtered, anonymized, and preprocessed to remove personally identifiable information (PII) and ensure privacy compliance [18], [71].

When selecting conversational data, researchers and practitioners must consider the following factors:

1) Data Quality [72]: Ensuring data quality is crucial to obtain reliable and accurate results. It involves assessing the credibility and authenticity of the data sources and applying appropriate filtering and cleaning techniques to remove noise, spam, or irrelevant content.
2) Data Relevance [45]: The selected conversational data should align with the research objectives or the specific application requirements. Consideration should be given

TABLE 4. Techniques for model compression and efficiency.

Technique	Description	Use Case	Examples
Knowledge Distillation [55]	Distilling knowledge from a larger, more complex model (teacher model) into a smaller model (student model) by training the student model to mimic the behavior and predictions of the teacher model. This technique helps reduce the model size while preserving performance.	Mobile devices, Edge computing, Knowledge Distillation in PyTorch	Hinton's Dark Knowledge, Neptune Project
Parameter Pruning [56]	Identifying and removing redundant or less important parameters in the model. Pruning techniques can be based on weight magnitudes, sensitivities, or iterative training methods. Pruned models have a reduced number of parameters, resulting in smaller model sizes and computational efficiency.	Image Classification Network Pruning, Pruning in PyTorch	LeCun's Optimal Brain Damage, NVIDIA TAO Toolkit
Quantization [57]	Reducing the precision of model weights and activations, typically from 32-bit floating-point representation to lower-bit fixed-point or integer representation. Quantization reduces memory requirements and accelerates computations by utilizing hardware optimizations for lower precision arithmetic.	Embedded systems, Mobile devices, Quantization in Machine Learning Models	Quantization Aware Training in TensorFlow
Knowledge Distillation with Pruning [58]	Combining knowledge distillation and parameter pruning techniques to compress models. The teacher model's knowledge is distilled into a smaller student model, and then pruning is applied to remove unnecessary parameters, further reducing the model size and computational requirements.	Improving the efficiency of deep neural networks in resource-constrained environments like mobile and edge devices.	PocketFlow, NVIDIA TAO Toolkit
Model Quantization with Pruning [59]	Similar to knowledge distillation with pruning, this technique combines parameter pruning with quantization. Pruning removes redundant parameters, and then quantization reduces the precision of remaining parameters, achieving even higher compression and efficiency.	Reducing the size and computational requirements of deep learning models without significant loss in performance [60].	TensorFlow Model Optimization Toolkit, Pruning preserving Quantization Aware Training (PQAT)

to the topics, domains, or target user groups to ensure that the collected data is relevant to the intended purpose.

3) Data Diversity [73]: It is important to have diverse conversational data that represents various demographic groups, cultural backgrounds, or linguistic variations. This helps in avoiding bias and obtaining a broader understanding of different conversational patterns and contexts.

To achieve these considerations, researchers and practitioners often employ data preprocessing techniques, such as tokenization and subword encoding, to transform the raw text into a suitable format for further analysis or training. Cleaning and filtering techniques may be applied to remove sensitive or offensive content and address ethical considerations.

A. DATA PREPROCESSING

Data preprocessing techniques play a crucial role in preparing and refining the training data for chatbot models [74]. These techniques ensure that the data is in a suitable format and quality for effective training. Here are some commonly used data preprocessing techniques for chatbot training:

- **Tokenization and Subword Encoding:** Tokenization involves breaking down sentences or texts into individual tokens or words. Subword encoding further subdivides tokens into smaller units, such as subwords or characters. These techniques help in representing the text in a structured format that can be processed by the large language model models [75].
- **Cleaning and Filtering Techniques:** Cleaning involves removing unnecessary elements from the text, such as punctuation marks, special characters, or HTML tags. Filtering techniques are employed to eliminate noise, spam, or irrelevant content from the dataset.These techniques help in enhancing the quality of the training data and reducing noise that may affect the large language model's performance [76].

TABLE 5. Comparison of GPT models.

Parameters	GPT	GPT-2	GPT-3	GPT-3.5	GPT-4
Model Size	Large	Very Large	Very Large	Extremely Large	Extremely Large
Number of Parameters	117 Million	1.5 Billion	175 Billion	175 Billion	1 Trillion
Training Data	Web Text	Web Text, Books	Web Text, Books	Web Text, Books	Web Text, Books
Context Window	512 tokens	1024 tokens	2048 tokens	4096 tokens	8000 tokens
Use Cases	Basic text generation	Improved text generation, language translation, summarization	Advanced text generation, language understanding, chatbots	Advanced text generation, code generation, language translation	Language translation, Customer service and Chatbot, Healthcare contract management
Fine-tuning	Not available	Available	Available	Available	Not available
Multilingual Support	Limited	Limited	Moderate	Moderate	High
Latency	Moderate	Moderate	High	High	High
Inference Speed	100 tokens/second	1000 tokens/second	10000 tokens/second	100000 tokens/second	100-200 tokens/second
Accuracy	83.1%	89.5%	94.4%	92.3%	97.1%
Fluency	Good	Good	Excellent	Excellent	Excellent
Creativity	Good	Good	Excellent	Excellent	Excellent

- **Balancing Dialogue Length and Dataset Size:** It is essential to balance the length of dialogues in the dataset to avoid biases and ensure that the model learns from dialogues of varying lengths. This can be achieved by randomly sampling dialogues or employing specific strategies to maintain a balanced distribution of dialogue lengths [77]. Additionally, considering the dataset size is important to ensure sufficient coverage and diversity in the training data.
- **Data Cleaning:** Data cleaning involves handling missing values, removing duplicates, and correcting inconsistencies in the dataset. Techniques such as imputation [78], dropping missing values [79], and handling outliers are used to ensure data quality and reliability for training the LLM's [80].
- **Data Transformation:** Data transformation includes scaling and normalizing numerical data to bring them to a similar range. This helps in avoiding biases towards features with larger values [81]. Additionally, encoding categorical variables using techniques like one-hot encoding or label encoding is used to represent categorical data in a format suitable for machine learning algorithms.
- **Feature Selection:** Feature selection involves selecting the most relevant and informative features from the dataset. It helps in reducing dimensionality, improving model performance, and reducing training time [82]. Techniques such as correlation analysis, backward/forward selection, and regularization methods like L1 and L2 regularization are used for feature selection [83].
- **Data Integration:** Data integration focuses on combining multiple data sources into a single, unified dataset. It involves handling inconsistencies in data formats, resolving naming conflicts, and merging datasets based on common attributes or keys. Data integration is crucial when working with diverse data sources and ensuring the compatibility of different datasets [84].
- **Text Preprocessing:** Text preprocessing techniques are specifically used for handling textual data. It includes tasks like tokenization, removing stop words, stemming or lemmatization to reduce words to their base form, and handling special characters or punctuation. Text preprocessing ensures that textual data is in a format suitable for natural language processing (NLP) tasks or text mining [85].

  The choice of techniques depends on the specific requirements of the dataset and the machine learning task at hand. Therefore, by applying these data preprocessing techniques, researchers and practitioners can effectively clean, structure, and prepare the training data for their large language models.

B. DATA AUGMENTATION AND SYNTHESIS TECHNIQUES

Data augmentation and synthesis techniques play a crucial role in enhancing the diversity and volume of conversational datasets, which are essential for training robust conversational models [86]. These techniques aim to reduce the dependence on labeled data and improve the performance of conversational systems in LLM's:

- **Conversational Data Augmentation (CODA):** CODA methods are used for semi-supervised abstractive conversation summarization [87]. These methods aim to reduce the dependence on labeled summaries by applying techniques such as random swapping/deletion to perturb the discourse relations inside conversations and dialogue-acts-guided insertion. CODA helps in expanding the available labeled dataset for training abstractive conversation summarization models.

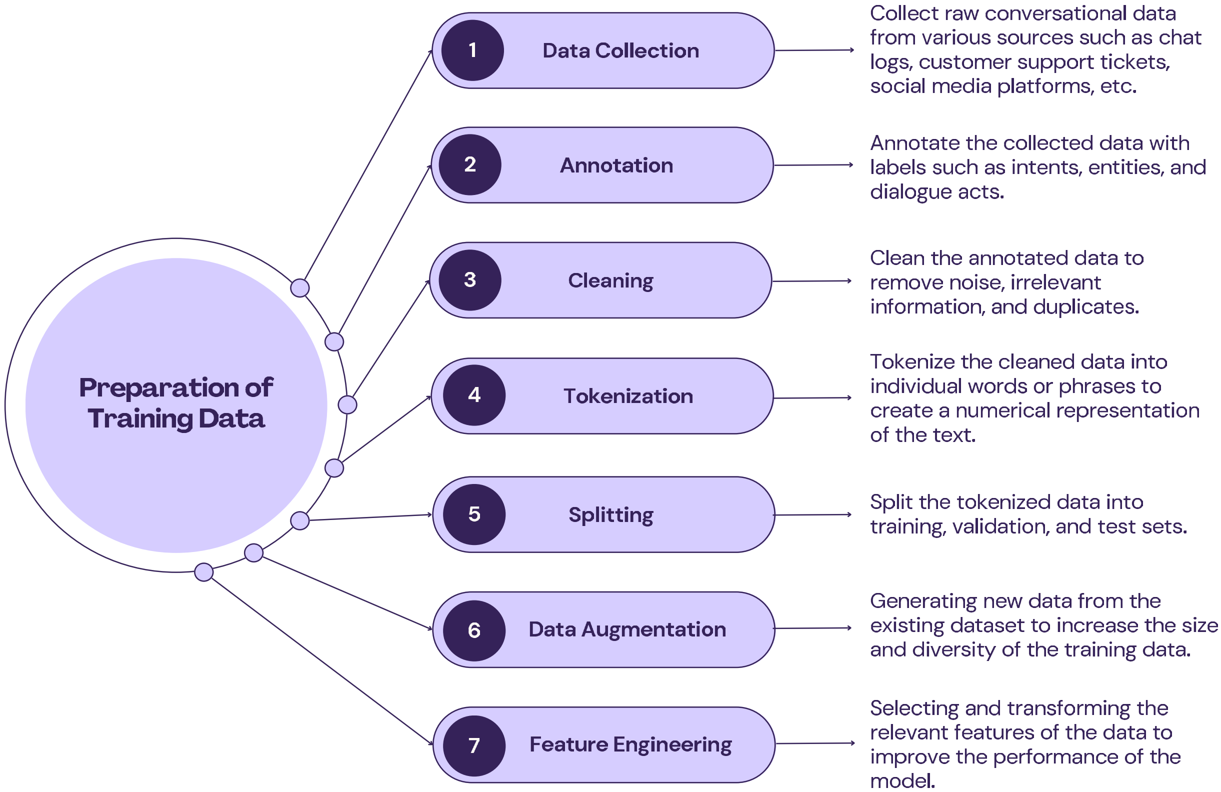


FIGURE 4. Training data pipeline.

- **Data Augmentation (DA) in Natural Language Processing:** Data augmentation is an effective strategy to alleviate data scarcity scenarios and improve LLM's performance [88]. Initially used in computer vision, DA has been introduced to natural language processing tasks as well. It involves creating modified copies of existing data by making minor changes or using deep learning techniques to generate new data points. DA can be applied to various machine learning applications where acquiring quality data is challenging, and it helps enhance model robustness and performance.
- **Simulated Dialogue Generation:** Simulated dialogue generation is a technique used for data expansion in conversational datasets. It involves generating synthetic dialogues using techniques such as rule-based approaches, reinforcement learning, or large language generative models like recurrent neural networks (RNNs) or transformer-based models. The simulation of new dialogues allows the dataset to be expanded, enabling a larger and more diverse training set for conversational models.
- **Text-to-Speech (TTS) Data Augmentation:** Text-to-Speech data augmentation has gained popularity in recent years. It involves synthesizing new audio samples from text using TTS models [89], [90]. The conditioning of TTS model on speaker-conditioning information and global style enables a broad variation coverage of data to be achieved. This approach can be suitable for augmenting conversational datasets with diverse speech samples.
- **Image Data Augmentation:** Image data augmentation is a widely used technique for increasing the diversity of training sets in computer vision tasks [91]. It involves applying random but realistic transformations to images, such as rotation, translation, scaling, captioning, flipping, and adding noise. These transformations help the model generalize better and improve its performance [92].
- **Audio Data Augmentation:** Data augmentation techniques are not limited to images, they can also be applied to audio data. For audio data augmentation, techniques such as noise injection can be used [93]. Adding Gaussian or random noise to the audio dataset helps improve the model's performance. By augmenting the audio data, the model can learn to handle various noise conditions and become more robust.

These approaches and techniques for data augmentation and synthesis in conversational datasets help overcome data scarcity, improve the LLM's performance, and enhance the quality and diversity of the training data. They provide valuable strategies for expanding datasets and addressing the challenges of training large language models.

C. ETHICAL CONSIDERATIONS AND PRIVACY

Ethical considerations and privacy in data usage, anonymization, and addressing sensitive or offensive content are crucial aspects of responsible data handling [94].

- **Legal and ethical considerations:** Balancing legal and ethical concerns is crucial in data usage. While legal standards are clear-cut, ethical standards can be subjective, emphasizing the importance of ethically using data to ensure privacy and societal benefits [95].
- **Anonymization and user consent:** Anonymization involves modifying data to safeguard individual privacy [96]. It's vital to have robust anonymization methods and to prioritize user consent, ensuring individuals are informed about their data's use.
- **Addressing sensitive or offensive content:** Addressing sensitive or offensive content requires careful handling considering legal and ethical aspects [97]. Organizations should adopt policies, moderation techniques, and community guidelines to manage such data responsibly.

It is essential for organizations and individuals to be aware of legal requirements, respect privacy rights, and uphold ethical standards when handling data [98]. Furthermore, by striking a balance between legal obligations, individual consent, and ethical considerations, we can promote responsible data usage that benefits both individuals and society as a whole [99].

In the next section, we explore the evolution of ChatGPT, tracing its development over time and examining the advancements that have propelled its capabilities in natural language processing (NLP).

VI. EVOLUTION OF CHATGPT

Conversational AI has witnessed significant advancements in recent years, enabling the development of more intelligent and human-like chatbots [100]. This section explores the evolution of Chat GPT, a state-of-the-art language model based on the GPT-3.5 architecture, and its contributions to the field of conversational AI [101]. Additionally, we discuss contextual understanding [102], coherence [103], long-range dependency modeling [104], improved handling of ambiguous queries or user intents [105], and a comparison with earlier chatbot models. Finally, we analyze the advantages and limitations of Chat GPT, including its strengths in generating natural and contextually relevant responses, challenges in controlling response quality and avoiding biases, as well as considerations regarding scalability and computational requirements [30].

A. DEVELOPMENT OF GPT FOR CHATBOT APPLICATIONS

Early conversational AI relied on rule-based systems and decision trees, which, while providing structure, lacked the flexibility to handle intricate conversations and couldn't learn from data [106], [107]. The emergence of Large Language Models and architectures like GPT, trained on extensive datasets, addressed these limitations. These models not only understand language nuances but also adapt to varied scenarios, recognizing long-range linguistic dependencies [52]. This advancement in chatbot technology, epitomized by GPT and other LLMs, has paved the way for more human-like interactions in areas such as customer support and virtual assistants, highlighting a transformative shift from rigid rule-based systems to dynamic, data-driven models [54], [108].

- **Contextual Understanding and Coherence:** One of the key strengths of GPT models in chatbot applications is their ability to understand and maintain context throughout a conversation. GPT models utilize contextual embeddings to capture relevant information from previous dialogue turns [109]. This contextual understanding allows the models to generate responses that consider the entire conversation history, making the interactions more fluid and coherent.
- **Long-Range Dependency Modeling:** Another significant advancement of GPT models in chatbot applications is their ability to model long-range dependencies within conversations [110]. Long-range dependencies refer to the relationships between words or concepts that are distant from each other in the conversation [111]. Capturing these dependencies is crucial for generating responses that address complex user queries or follow nuanced conversation threads.
  GPT models leverage their self-attention mechanism, which allows them to assign different weights to different parts of the dialogue history [82]. This mechanism enables the models to identify and remember important information from earlier dialogue turns, even when there are multiple intervening interactions.
- **Improved Handling of Ambiguous Queries or User Intents:** GPT models have demonstrated improved capabilities in handling ambiguous queries or user intents [112]. Ambiguity in user inputs can arise from various sources, such as homonyms, vague language, or incomplete sentences. GPT models benefit from their extensive training on diverse text data, allowing them to learn contextual cues and disambiguate user queries effectively [113]. When faced with ambiguous queries, GPT models consider the context of the conversation and leverage the knowledge learned from the training data to generate responses that are most likely to align with the user's intended meaning. This enhanced handling of ambiguous queries or intents contributes to more accurate and contextually appropriate responses, providing users with a more satisfying chatbot experience.

B. ADVANCEMENTS IN CHATBOT MODELS

In the context of comparing ChatGPT with earlier chatbot models, several theories and frameworks can be discussed to highlight the advancements and differences.

- **Rule-based systems and decision trees:** Early chatbot models leaned heavily on rule-based systems and decision trees. Rule-based systems utilized predefined

IF-THEN rules, triggering specific responses based on set conditions [114], [115]. In contrast, decision trees used traversal algorithms, often guided by methods like ID3 or C4.5, to navigate the conversation based on input and node-associated conditions [116], [117]. Despite their structured approach, these systems struggled with intricate, open-ended dialogues [118]. The introduction of Large Language Models heralded a significant shift, offering a more dynamic and adaptive solution, better equipped to handle the complexities of human language and conversation [119].

– **Retrieval-based approach:** Chatbots traditionally use retrieval-based and rule-based approaches [120]. Retrieval-based chatbots match user queries to pre-set responses using similarity measures, while rule-based ones generate responses based on predefined rules [121]. Two primary similarity measures are:
  a) Cosine Similarity: It measures the cosine angle between two vectors, helping compare a user's query vector with existing response vectors [122].
  b) Jaccard Similarity: It compares two sets by evaluating the ratio of their intersection to their union. It's useful for comparing user query sets to existing response sets [123].
– **Improved Handling of Ambiguous Queries or User Intents:**
  Sequence-to-sequence models, coupled with attention mechanisms, are widely used in chatbot development [124]. These models enable the generation of coherent and contextually relevant responses by capturing the dependencies between input and output sequences [125]. Sequence-to-sequence models consist of two components: an encoder and a decoder.
  a) Encoder: The encoder takes the input sequence (user query) and transforms it into a fixed-length vector representation called the "context vector" or "hidden state" [126]. This process captures the semantic meaning and context of the input sequence. The encoder can be implemented using various architectures, such as Recurrent Neural Networks (RNNs) or transformer models [127], [128].
  b) Decoder: The decoder takes the context vector and generates the output sequence (chatbot response) [129], [130]. It does so by iteratively predicting the next token in the output sequence based on the previous tokens generated. The decoder can also employ recurrent neural networks or transformer models [131].
  c) Attention Mechanism: Attention mechanisms improve the performance of sequence-to-sequence models by allowing the model to focus on different parts of the input sequence during the decoding process [132], [133]. Attention mechanisms assign weights to different positions in the input sequence, indicating their relevance to the current decoding step. These weights are used to compute a weighted sum of the encoder's hidden states, providing the decoder with the necessary context information [134], [135].

  The integration of sequence-to-sequence models and attention mechanisms enhances the chatbot's ability to capture long-range dependencies, maintain context throughout the conversation, and generate coherent and contextually appropriate responses.

VII. ADVANTAGES AND LIMITATIONS OF CHAT GPT

Chat GPT, an advanced large language AI model, offers several advantages and brings new possibilities to the world of chatbot applications [101]. Its strengths lie in generating natural and contextually relevant responses, enhancing user engagement and language flexibility. However, like any technology, Chat GPT also has its limitations. Controlling response quality, addressing biases, and managing scalability and computational requirements present challenges that need to be carefully considered. Understanding both the advantages and limitations of Chat GPT is crucial for effectively leveraging its capabilities and designing responsible and efficient chatbot systems [136].

– **Advantages:**
  - ***Generating Natural and Contextually Relevant Responses:***
    Chat GPT excels in generating responses that are natural-sounding and contextually relevant. It has the ability to understand and incorporate context from previous dialogue turns, resulting in more coherent and meaningful conversations.
  - ***Language Understanding and Fluency:***
    Chat GPT is trained on vast amounts of text data, allowing it to capture the complexities of human language. This enables it to generate responses that are syntactically and semantically correct, leading to more fluent and coherent interactions with users [137].
  - ***Adaptability to Different Domains:***
    Chat GPT can be fine-tuned on specific datasets or domains, making it adaptable to different applications. This flexibility allows it to provide more accurate and domain-specific responses, enhancing its usefulness in various industries and domains [138].
  - ***Improved User Engagement:***
    Chat GPT's ability to generate natural and contextually relevant responses enhances user engagement. It creates more interactive and dynamic conversations, increasing user satisfaction and fostering a positive user experience [139].
  - ***Language Flexibility:***
    Chat GPT is capable of understanding and generating responses in multiple languages. It can handle diverse linguistic inputs, catering to users from different language backgrounds and facilitating cross-cultural communication [140].

TABLE 6. Comparison of earlier Chatbot models with ChatGPT.

Model	Approach	Advantages	Limitations	ChatGPT's Contributions
ELIZA	Rule-based	Simple and intuitive dialogue patterns.	Limited to predefined rules and lacks understanding of context.	ChatGPT surpasses ELIZA by offering more sophisticated and contextually aware responses through machine learning.
ALICE	Pattern-matching	Can handle a wide range of topics.	Relies heavily on pattern matching, lacks deeper understanding.	ChatGPT surpasses ALICE by providing more coherent and contextually relevant responses based on a broader understanding of language.
Jabber-wacky	Learning from conversations	Learns from user input and adapts over time.	Tends to generate nonsensical or irrelevant responses.	ChatGPT outperforms Jabberwacky by offering more coherent and contextually appropriate responses with a better understanding of conversational nuances.
Mitsuku	Rule-based with learning capabilities	Good at engaging in casual conversations.	Limited understanding of complex queries and lacks deep knowledge.	ChatGPT surpasses Mitsuku by offering a broader understanding of language, handling more complex queries, and providing more informative responses.
ALICE AI Foundation	AIML-based	Supports multiple languages.	Prone to generating generic or scripted responses.	ChatGPT outperforms ALICE AI Foundation by generating more diverse and contextually relevant responses, avoiding the tendency for generic or scripted outputs.
Cleverbot	Machine learning from user input	Can hold longer and more coherent conversations.	Tends to give inconsistent responses and may not understand nuanced queries.	ChatGPT surpasses Cleverbot by offering a higher level of coherence and understanding, providing more accurate and contextually relevant responses.

- ***Knowledge Base Integration:***
  Chat GPT can be integrated with external knowledge bases or information sources. This allows it to provide factual and informative responses, leveraging the vast amount of knowledge available on the internet or specific domains.

– **Limitations:**
While ChatGPT showcases impressive conversational abilities, it is essential to recognize its potential biases, the variability in output quality, and the ethical considerations entailed in its deployment.

- ***Controlling Response Quality:***
  While Chat GPT is proficient in generating responses, controlling the quality of those responses can be challenging [141]. It may occasionally produce inaccurate or irrelevant answers, requiring robust mechanisms to ensure response correctness and reliability.
- ***Bias and Ethical Concerns:***
  Chat GPT can inadvertently exhibit biases present in the training data, potentially leading to biased or discriminatory responses [142]. Addressing bias and ensuring fairness in chatbot interactions is an ongoing challenge that requires careful monitoring, data curation, and bias mitigation techniques [143].
- ***Scalability and Computational Requirements:***
  Chat GPT's large-scale architecture and training process require substantial computational resources [144]. Deploying and scaling up Chat GPT systems can be computationally intensive and expensive, limiting its accessibility for organizations with limited resources.
- ***Lack of Common Sense Reasoning:***
  Chat GPT may struggle with common sense reasoning and understanding subtle nuances or implicit information [145]. It can sometimes generate

responses that are technically correct but lack real-world common sense, leading to occasional misunderstandings or incorrect answers. It also generates inconsistent responses.

- ***Contextual Inconsistencies:***
  While Chat GPT can capture context from previous dialogue turns, it may still struggle with long-term context retention [146]. In complex conversations or discussions spanning multiple topics, it may not maintain consistent context awareness throughout the dialogue.
- ***Interpretation of Ambiguous Queries:***
  Ambiguous queries or user intents can pose challenges for Chat GPT. It may struggle to disambiguate unclear or vague inputs, leading to potential misunderstandings or inaccurate responses [147]. Handling ambiguity effectively remains an ongoing research challenge.
- ***Unable to provide up-to-date information yet:***
  Chat GPT has a knowledge cut-off date of 2021(so far). It may be limited to non-current data, which, in turn, can cause its users to suffer from knowledge gaps, or worse, false information when the tool misinterprets prompts, or turns dishonest, if any how possible.

In the next section, we discuss about the training process of ChatGPT, exploring the techniques used to train this powerful language model.

VIII. TRAINING PROCESS OF CHAT GPT

In this section, we present the detailed structure of a very famous large language model, ChatGPT, an artificial intelligence model designed for conversational applications. ChatGPT leverages the GPT-3.5 architecture developed by OpenAI and incorporates several key components to enable effective and coherent chat-based interactions. The structure of ChatGPT can be described as follows:

- **Preprocessing:**
  - Tokenization: The input text is tokenized into a sequence of tokens to facilitate further processing.
  - Special Tokens: Special tokens such as [CLS] [148] and [SEP] [149] are added to mark the beginning and end of the input sequence and to handle variable-length conversations.
  - Input Encoding: The tokenized input is then encoded into numerical representations that can be processed by the neural network.
- **GPT-3.5 Architecture:**
  - Transformer Encoder: ChatGPT utilizes a multi-layer Transformer encoder architecture. Each layer consists of a self-attention mechanism and position-wise feed-forward neural networks. The encoder processes the input tokens and captures the contextual dependencies within the conversation.
  - Attention Mechanism: Self-attention allows ChatGPT to attend to different parts of the conversation, giving higher weightage to relevant tokens while generating responses.
  - Positional Encoding: Positional encoding is added to the input tokens to provide information about their relative positions in the conversation, enabling the model to understand the sequential nature of the text.
- **Training Process:**
  - Pretraining: ChatGPT is pretrained on a large corpus of publicly available text data, enabling it to learn the statistical properties of language.
  - Fine-tuning: After pretraining, ChatGPT is fine-tuned on a specific conversational dataset. The fine-tuning process involves exposing the model to conversational examples and optimizing its parameters to generate appropriate and contextually relevant responses.
- **Inference and Response Generation:**
  - Decoding Strategy: During inference, ChatGPT utilizes a decoding strategy such as beam search or top-k sampling to generate diverse and plausible responses.
  - Temperature Parameter: A temperature parameter is used to control the randomness of the generated responses. Higher values produce more random responses, while lower values yield more focused and deterministic outputs.
  - Context Window: ChatGPT incorporates a context window mechanism to limit the length of the input conversation that is considered when generating responses, ensuring computational efficiency.
- **Evaluation Metrics:**
  - Coherence: The coherence of ChatGPT's responses can be evaluated using metrics such as perplexity or language modeling scores.
  - Context Sensitivity: The model's ability to generate contextually relevant responses can be assessed through human evaluation or by comparing the generated responses with ground truth responses in a conversational dataset.

ChatGPT's structure encompasses preprocessing steps, the GPT-3.5 architecture, training process, inference and response generation strategies, evaluation metrics, and considerations for limitations and future improvements. This comprehensive structure enables ChatGPT to generate human-like responses in conversational applications. It is essential to note that while ChatGPT provides insights based on a vast amount of data, its knowledge is limited to its last training cut-off in 2021. Thus, it may not be aware of developments or information post-2021.

In the next section, we delve into the evaluation metrics employed to assess the performance and effectiveness of ChatGPT and similar language models. Accurately evaluating these models is crucial for understanding their capabilities and identifying areas for improvement.

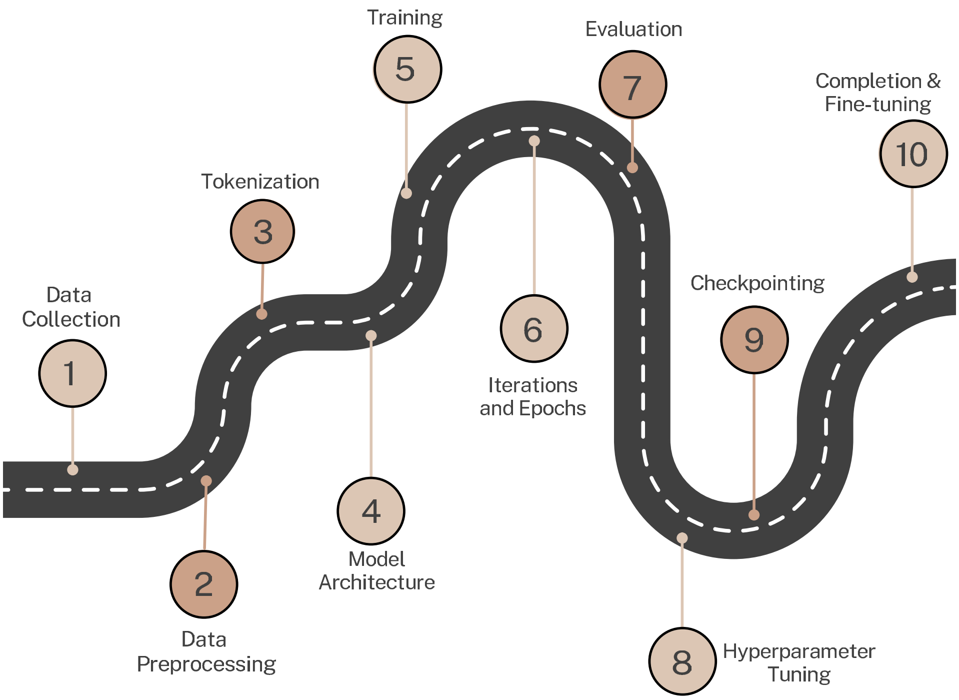


FIGURE 5. **Step-by-step training process of chat-GPT.**

IX. EVALUATION METRICS

Evaluating the performance of conversational AI systems, including ChatGPT, requires the use of appropriate evaluation metrics [150].

1) Precision: Precision measures the accuracy of a model by calculating the ratio of correctly identified positives (true positives) to all identified positives. It indicates how many of the predicted classes are correctly labeled.
2) Recall: Recall measures the model's ability to predict actual positive classes by calculating the ratio of predicted true positives to the total number of instances that should have been tagged [151]. It reveals how many of the predicted classes are correct.
3) F1 Score: The F1 score is a combined metric that considers both precision and recall. It provides a balance between these two metrics and is calculated as harmonic mean of recall and precision [152].
4) Confusion Matrix: A confusion matrix is an N x N matrix that compares the expected labels with the ones predicted by the model. It gives a holistic view of how well the model is performing and helps identify specific errors or ambiguities in classification.

A. USER SATISFACTION AND ENGAGEMENT METRICS FOR ASSESSING CHATBOT PERFORMANCE

User Satisfaction and Engagement Metrics play a crucial role in evaluating the performance of chatbots [153]. These metrics focus on measuring the overall user experience, satisfaction, and engagement with the chatbot. Here are some key aspects to consider:

– **Customer Satisfaction:** One metric to assess chatbot performance is customer satisfaction. It measures the percentage of users who find the chatbot helpful or a score based on common customer satisfaction metrics [16]. Directly asking users about their satisfaction with the chatbot can provide reliable insights.
– **User Feedback and Surveys:** Gathering feedback from users through surveys or structured questionnaires can provide valuable insights into their experience with the chatbot [154]. These feedback mechanisms allow users to express their satisfaction levels, highlight areas for improvement, and provide suggestions to enhance the chatbot's performance.
– **Response Rates and Engagement:** Monitoring response rates and user engagement metrics can indicate the effectiveness of the chatbot. Metrics such as average response time, conversation length, or number of interactions per user session can reflect user engagement [155]. Higher engagement levels generally indicate a positive user experience and effective interaction with the chatbot.
– **Issue Resolution:** Tracking the chatbot's ability to successfully resolve user issues can be an important metric. It measures the number of issues or queries the chatbot was able to resolve without the need for human intervention [156]. Higher resolution rates indicate the chatbot's efficiency and effectiveness in addressing user concerns.
– **Conversion and Leads:** For chatbots involved in sales or lead generation, metrics such as the number or quality of leads generated can be essential. Assessing the

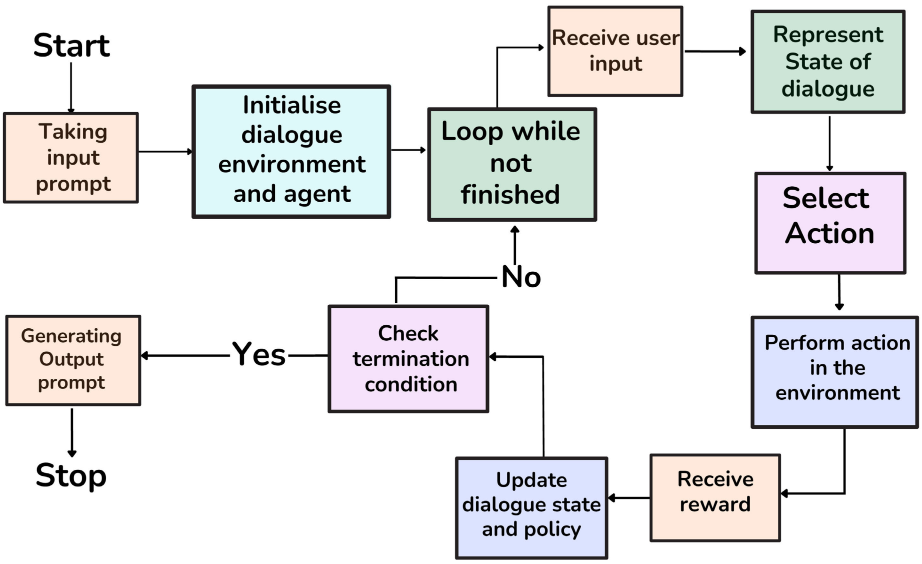


FIGURE 6. Reinforcement learning for dialogue management flowchart.

chatbot's impact on lead generation and its contribution to the sales funnel can help evaluate its performance and effectiveness in driving business outcomes.

- **Cost Efficiency:** Evaluating the cost per issue resolved or cost per interaction can provide insights into the chatbot's efficiency and cost-effectiveness compared to other customer service channels. Cost efficiency metrics help businesses assess the value and Return On Investment (ROI) of implementing chatbot solutions.

B. PERFORMANCE AGAINST OTHER STATE-OF-THE-ART MODELS FOR COMPARATIVE EVALUATION OF CHAT GPT

Comparing the performance of Chat GPT with other state-of-the-art chatbot models is a crucial aspect of evaluating its capabilities and advancements in natural language understanding and generation [152].

- **Evaluation on Benchmark Datasets:** Comparative evaluation often involves using benchmark datasets and evaluation frameworks. These standardized datasets contain a range of conversation scenarios and tasks, allowing fair and consistent comparison between different chatbot models [157].
- **Comparison with Retrieval-Based Approaches:** One approach to evaluate Chat GPT is to compare it with retrieval-based models. Retrieval-based chatbots retrieve pre-defined responses from a knowledge base or a set of predefined responses based on keyword matching or similarity measures [158].
- **Comparison with Rule-Based Approaches:** Rule-based chatbots rely on a predefined set of rules or patterns to generate responses [159]. Comparing Chat GPT with rule-based models helps evaluate its ability to generate more natural and human-like responses by leveraging its language modeling capabilities and contextual understanding.
- **Evaluation Metrics and Tasks:** Comparative evaluation involves assessing various metrics and tasks. These can include measuring response coherence, grammaticality, relevance to user queries, and overall conversational quality [153]. Metrics such as Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and perplexity can be employed to quantitatively evaluate the model's performance.
- **Evaluation on GLUE Benchmark:** The General Language Understanding Evaluation (GLUE) benchmark is commonly used to evaluate the understanding ability of large language models, including chatbot models [160]. Comparing Chat GPT's performance on the GLUE benchmark with other fine-tuned BERT-style models provides insights into its language comprehension and contextual understanding abilities [161].
- **Comparative Analysis of Features and Limitations:** A comprehensive evaluation involves a comparative analysis of LLM's features, limitations, and capabilities against other state-of-the-art models. This analysis helps identify the unique strengths and weaknesses of LLM's like Chat GPT and provides insights into its potential applications and areas for improvement.

In the next section, we explore the societal impact of ChatGPT and its implications on various aspects of our lives.

X. SOCIETAL IMPACT

The remarkable capabilities of ChatGPT have been harnessed by individuals and organizations across different demographics, leading to numerous beneficial outcomes.

1) **Assisting Customer Support:** ChatGPT can improve customer support services by providing accurate and timely responses, enhancing user experiences and satisfaction.

2) **Language Translation and Understanding:** ChatGPT's language translation capabilities enable effective communication across language barriers, fostering inclusivity and facilitating global interactions [162].
3) **Content Generation and Summarization:** ChatGPT aids content creators in generating ideas, refining drafts, and summarizing lengthy documents, improving productivity and creativity.
4) **Education and Learning Support:** ChatGPT can assist students by providing explanations, clarifications, and learning resources, augmenting educational experiences and promoting knowledge acquisition.
5) **Accessibility and Inclusion:** ChatGPT's multilingual support and conversational abilities make information more accessible, bridging communication gaps and empowering individuals with diverse linguistic backgrounds.
6) **Support for Professional Tasks:** ChatGPT assists professionals in various domains, such as providing technical guidance, answering common queries, and aiding in decision-making processes, leading to increased efficiency and effectiveness.

The remarkable capabilities of ChatGPT have not gone unnoticed by individuals seeking personal gain. From students to elderly people, various demographics have been implicated in misusing and exploiting ChatGPT for their own benefit. This section delves into how different groups have employed ChatGPT inappropriately, highlighting the consequences and implications of such actions.

1) **Students:**
– ***Academic Dishonesty:*** Some students have resorted to using ChatGPT to generate plagiarized essays, assignments, or academic papers. By inputting questions or prompts, they exploit the model to produce seemingly original content. This undermines the learning process, intellectual honesty, and the integrity of educational institutions [163].
– ***Cheating in Examinations:*** ChatGPT's ability to generate real-time responses has led to its misuse during examinations. Students attempt to obtain answers to exam questions by surreptitiously feeding prompts to the model via electronic devices. This dishonest practice compromises the fairness and credibility of assessment systems [164].
2) **Professionals:**
– ***Automated Content Creation:*** Content creators and writers may misuse ChatGPT to automate content production. While the technology can aid in generating ideas and inspiration, relying solely on ChatGPT-generated content without proper attribution or editing is unethical. It devalues the creative efforts of individuals and can lead to a flood of low-quality, generic content [165].
– ***Fake Reviews and Testimonials:*** Some professionals resort to using ChatGPT to generate fabricated positive reviews or testimonials to manipulate public perception and bolster their reputations [166]. This deceptive practice undermines trust in online reviews, compromises consumer decision-making, and distorts market dynamics [167].
3) **Scammers and Fraudsters:**
– ***Phishing and Social Engineering:*** Malicious actors exploit ChatGPT to craft persuasive messages for phishing attacks or social engineering schemes [168]. By creating seemingly authentic and personalized messages, scammers aim to deceive individuals into sharing sensitive information, making financial transactions, or engaging in harmful activities.
– ***Advanced Fraud Techniques:*** ChatGPT's conversational abilities can be harnessed by fraudsters to develop sophisticated scams. They manipulate individuals into providing personal information, initiating financial transactions, or falling victim to deceptive investment schemes. This form of exploitation can result in financial loss and emotional distress for unsuspecting individuals [169].
4) **Elderly Individuals:**
– ***Elder Abuse:*** The elderly, who may be less familiar with emerging technologies, can be targeted for various forms of exploitation. Misusers of ChatGPT may pose as helpful individuals or organizations, engaging in conversations to deceive and manipulate the elderly into divulging personal information, sharing financial details, or falling victim to scams.
– ***Psychological Manipulation:*** Some unscrupulous individuals exploit ChatGPT's conversational capabilities to engage in emotional manipulation and coercion. They may attempt to establish false relationships, exploit vulnerabilities, or extort money from elderly individuals who may be seeking companionship or support.

Misuse of ChatGPT spans from students seeking shortcuts to scammers targeting the vulnerable [170]. This misuse damages trust and societal integrity. Addressing this requires promoting ethical AI, improving digital literacy, and implementing safeguards against misuse [171]. Also, education, awareness campaigns, and legal actions are essential to counteract these practices [163].

XI. CHALLENGES AND OPEN ISSUES

ChatGPT, a conversational AI, faces issues in cybersecurity, understanding, ethics, context retention, overconfidence, data dependency, and user well-being [172], [173]. It can produce biased content, lack context, and show overconfidence in wrong answers. Emphasizing user safety, mental health, and privacy is vital [174]. Research continues to address these challenges and improve ChatGPT's safety and proficiency.

– **Vulnerable to cyber-crime:** ChatGPT introduces cybersecurity risks, enabling hackers to exploit defense vulnerabilities [175]. With rising cyberattacks and data breaches, there's an urgent need to address these security threats [176]. Malicious users can use ChatGPT for deceptive communications, including phishing [176]. As ChatGPT advances, it might face adversarial attacks targeting its weaknesses. To counter these risks, multi-layered defense systems, detection mechanisms, and strict authentication are vital [177]. Continuous research and AI-cybersecurity [178] collaboration are essential for addressing these vulnerabilities.

- **Limitations in Natural Language Understanding:** ChatGPT has limitations in understanding and generating natural language [179]. It can produce inaccurate responses, especially for complex or ambiguous queries, due to the challenges of grasping human language nuances [180]. Efforts to address these constraints involve diverse training data, improved contextual reasoning, and enhancing the model's handling of ambiguity.
- **Ethical and Social Implications:** ChatGPT's use can lead to biased content and potential misuse [181]. If trained on skewed data, it might reinforce societal biases, leading to discrimination. Its human-like generation can be misused for misinformation. To address this, transparency, diverse data, bias checks, and AI explanations are crucial. Following established AI guidelines and standards ensures alignment with societal values. Ongoing research, public discourse, and policy considerations are key to responsibly shaping AI's future [182].
- **Lack of Context and Memory:** ChatGPT's limited memory and context retention hinder sustained coherent conversations [183]. It often forgets past information, leading to inconsistencies or repetition since it views each input independently, without a conversation history [184]. To address this, researchers explore memory-augmented architectures and attention mechanisms [31]. Users can help by providing context or reminders.
- **Overconfidence and Incorrect Answers:** ChatGPT can be overconfident, even when offering incorrect answers [185]. This can mislead users into trusting inaccurate information. Its confidence arises from diverse training data, which includes both accurate and erroneous content. As ChatGPT can't verify its answers, users should corroborate with trusted sources and fact-check [186]. Efforts are underway to improve model confidence calibration. Responsible use of ChatGPT involves understanding its limitations and combining its insights with critical evaluation [187].
- **Dependency on Training Data:** ChatGPT's reliance on training data can lead to biases and inaccuracies in its outputs [188]. If the data is skewed or flawed, the model's responses can be affected [189]. Addressing this requires careful data curation and bias-detection techniques. Users should be aware of these biases and verify ChatGPT's responses with trusted sources.
- **User Safety and Well-being:** ChatGPT's impact on user well-being, especially vulnerable individuals, raises safety, mental health, and privacy concerns [190]. Addressing this requires designing ChatGPT with user safety and privacy in mind, curating training data to avoid harmful content, and ensuring secure data handling. Clear guidelines, content moderation, and emotional support resources are crucial. Collaborations between AI developers and mental health experts can help mitigate risks. By prioritizing user well-being, we ensure responsible use of AI like ChatGPT [191].

XII. CONCLUSION

In conclusion, Conversational AI and various Large Language Models like ChatGPT have emerged as powerful technologies that have revolutionized the field of human-computer interaction. This survey paper aimed to explore and analyze the current state of Conversational AI and specifically focus on the capabilities and limitations of ChatGPT, a prominent example of such technology. Throughout this paper, we examined the fundamental concepts and components of Conversational AI, including natural language understanding, dialogue management, and natural language generation. We also delved into the underlying techniques used in training and fine-tuning ChatGPT, highlighting its reliance on deep learning, reinforcement learning, and large-scale pre-training. Moreover, we discussed the significant advancements achieved in Conversational AI, with ChatGPT showcasing impressive performance in generating coherent and contextually relevant responses. The ability of ChatGPT to engage users in human-like conversations has garnered widespread attention and applications across various domains, including customer support, virtual assistants, and content generation.

REFERENCES

[1] V. Gaur and N. Saunshi, "Symbolic math reasoning with language models," in *Proc. IEEE MIT Undergraduate Res. Technol. Conf. (URTC)*, Sep. 2022, pp. 1–5.

[2] K. Chen, T. Zhao, M. Yang, L. Liu, A. Tamura, R. Wang, M. Utiyama, and E. Sumita, "A neural approach to source dependence based context model for statistical machine translation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 266–280, Feb. 2018.

[3] A. Mishra, A. Anand, and P. Guha, "Dual attention and question categorization-based visual question answering," *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 81–91, Feb. 2023.

[4] M. Yang, C. Li, Y. Shen, Q. Wu, Z. Zhao, and X. Chen, "Hierarchical human-like deep neural networks for abstractive text summarization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2744–2757, Jun. 2021.

[5] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, S. Mishra, and A. Abraham, "AI-based conversational agents: A scoping review from technologies to future directions," *IEEE Access*, vol. 10, pp. 92337–92356, 2022.

[6] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," 2015, *arXiv:1506.06714*.

[7] G. Bansal, V. Chamola, P. Narang, S. Kumar, and S. Raman, "Deep3DSCan: Deep residual network and morphological descriptor based framework forlung cancer classification and 3D segmentation," *IET Image Process.*, vol. 14, no. 7, pp. 1240–1247, May 2020.

[8] J. Xie, F. Zhu, M. Huang, N. Xiong, S. Huang, and W. Xiong, "Unsupervised learning of paragraph embeddings for context-aware recommendation," *IEEE Access*, vol. 7, pp. 43100–43109, 2019.

[9] H. Chen, L. F. Pieptea, and J. Ding, "Construction and evaluation of a high-quality corpus for legal intelligence using semiautomated approaches," *IEEE Trans. Rel.*, vol. 71, no. 2, pp. 657–673, Jun. 2022.

[10] A. Madasu and S. Srivastava, "What do large language models learn beyond language?" 2022, *arXiv:2210.12302*.

[11] X. Liu, G. Zhou, M. Kong, Z. Yin, X. Li, L. Yin, and W. Zheng, "Developing multi-labelled corpus of Twitter short texts: A semi-automatic method," *Systems*, vol. 11, no. 8, p. 390, Aug. 2023.

[12] H. Strobelt, J. Kinley, R. Krueger, J. Beyer, H. Pfister, and A. M. Rush, "GenNI: Human-AI collaboration for data-backed text generation," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 1106–1116, Jan. 2022.

[13] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Machine learning techniques for biomedical natural language processing: A comprehensive review," *IEEE Access*, vol. 9, pp. 140628–140653, 2021.

[14] P. Sitkrongwong, A. Takasu, and S. Maneeroj, "Context-aware user and item representations based on unsupervised context extraction from reviews," *IEEE Access*, vol. 8, pp. 87094–87114, 2020.

[15] A. Ramaditiya, S. Rahmatia, A. Munawar, and O. N. Samijayani, "Implementation chatbot whatsapp using Python programming for broadcast and reply message automatically," in *Proc. Int. Symp. Electron. Smart Devices (ISESD)*, Jun. 2021, pp. 1–4.

[16] Z. Liu, C. Long, X. Lu, Z. Hu, J. Zhang, and Y. Wang, "Which channel to ask my question? Personalized customer service request stream routing using deep reinforcement learning," *IEEE Access*, vol. 7, pp. 107744–107756, 2019.

[17] S. Floyd, "Identifying variables that improve communication with bots," in *Proc. IEEE Int. Symp. Meas. Control Robot. (ISMCR)*, Sep. 2019, pp. C3-3-1–C3-3-6.

[18] L. M. Sánchez-Adame, S. Mendoza, J. Urquiza, J. Rodríguez, and A. Meneses-Viveros, "Towards a set of heuristics for evaluating chatbots," *IEEE Latin Amer. Trans.*, vol. 19, no. 12, pp. 2037–2045, Dec. 2021.

[19] M. Ochs, K. Narasimhan, and M. Mezini, "Evaluating and improving transformers pre-trained on ASTs for code completion," in *Proc. IEEE Int. Conf. Softw. Anal., Evol. Reeng. (SANER)*, Mar. 2023, pp. 834–844.

[20] R. Ren, S. Pérez-soler, J. W. Castro, O. Dieste, and S. T. Acuña, "Using the SOCIO chatbot for UML modeling: A second family of experiments on usability in academic settings," *IEEE Access*, vol. 10, pp. 130542–130562, 2022.

[21] V. Bilgram and F. Laarmann, "Accelerating innovation with generative AI: AI-augmented digital prototyping and innovation methods," *IEEE Eng. Manag. Rev.*, vol. 51, no. 2, pp. 18–25, 2023, doi: 10.1109/EMR.2023.3272799.

[22] G. Daniel, J. Cabot, L. Deruelle, and M. Derras, "Xatkit: A multimodal low-code chatbot development framework," *IEEE Access*, vol. 8, pp. 15332–15346, 2020.

[23] F.-Y. Wang, J. Yang, X. Wang, J. Li, and Q.-L. Han, "Chat with ChatGPT on Industry 5.0: Learning and decision-making for intelligent industries," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 831–834, Apr. 2023.

[24] F.-Y. Wang, Q. Miao, X. Li, X. Wang, and Y. Lin, "What does ChatGPT say: The DAO from algorithmic intelligence to linguistic intelligence," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 3, pp. 575–579, Mar. 2023.

[25] F.-Y. Wang, J. Li, R. Qin, J. Zhu, H. Mo, and B. Hu, "ChatGPT for computational social systems: From conversational applications to human-oriented operating systems," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 2, pp. 414–425, Apr. 2023.

[26] G. Hurlburt, "What if ethics got in the way of generative AI?" *IT Prof.*, vol. 25, no. 2, pp. 4–6, Mar. 2023.

[27] L. Zhang, Z. Sun, J. Zhang, Y. Wu, and Y. Xia, "Conversation-based adaptive relational translation method for next POI recommendation with uncertain check-ins," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[28] H. Honda and M. Hagiwara, "Question answering systems with deep learning-based symbolic processing," *IEEE Access*, vol. 7, pp. 152368–152378, 2019.

[29] P. Maddigan and T. Susnjak, "Chat2VIS: Generating data visualizations via natural language using ChatGPT, Codex and GPT-3 large language models," *IEEE Access*, vol. 11, pp. 45181–45193, 2023.

[30] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1122–1136, May 2023.

[31] S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu, and W. Zheng, "The multi-modal fusion in visual question answering: A review of attention mechanisms," *PeerJ Comput. Sci.*, vol. 9, p. e1400, May 2023.

[32] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. Riggs, M. Esmaeili, R. Majdabadkohne, and M. Pasehvar, "ChatGPT: Applications, opportunities, and threats," in *Proc. SIEDS*, Apr. 2023, pp. 274–279.

[33] N. Chowdhury, O. A. Awais, and S. Aktar, "Improving customer care with ChatGPT: A case study," Tech. Rep., Mar. 2023.

[34] Y. Gao, R. Wang, and F. Hou, "How to design translation prompts for ChatGPT: An empirical study," 2023, *arXiv:2304.02182*.

[35] H. Zhang, X. Liu, and J. Zhang, "Extractive summarization via ChatGPT for faithful summary generation," 2023, *arXiv:2304.04193*.

[36] P. S. V. Reddy, T. P. Kalki, P. Roshini, and S. Navaneethan, "Varoka-chatbot: An artificial intelligence based desktop partner," in *Proc. Int. Conf. Artif. Intell. Knowl. Discovery Concurrent Eng. (ICECONF)*, Jan. 2023, pp. 1–6.

[37] X. Liu, T. Shi, G. Zhou, M. Liu, Z. Yin, L. Yin, and W. Zheng, "Emotion classification for short texts: An improved multi-label method," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 1–9, Jun. 2023.

[38] L. Li, P. Wang, X. Zheng, Q. Xie, X. Tao, and J. D. Velásquez, "Dual-interactive fusion for code-mixed deep representation learning in tag recommendation," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101862.

[39] A. Shafeeg, I. Shazhaev, D. Mihaylov, A. Tularov, and I. Shazhaev, "Voice assistant integrated with Chat GPT," *Indonesian J. Comput. Sci.*, vol. 12, no. 1, pp. 22–31, Feb. 2023.

[40] Q. Ni, J. Guo, W. Wu, H. Wang, and J. Wu, "Continuous influence-based community partition for social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 3, pp. 1187–1197, May 2022.

[41] S. García-Méndez, F. De Arriba-Pérez, F. J. González-Castaño, J. A. Regueiro-Janeiro, and F. Gil-Castiñeira, "Entertainment chatbot for the digital inclusion of elderly people without abstraction capabilities," *IEEE Access*, vol. 9, pp. 75878–75891, 2021.

[42] W. Nie, Y. Bao, Y. Zhao, and A. Liu, "Long dialogue emotion detection based on commonsense knowledge graph guidance," *IEEE Trans. Multimedia*, 2023.

[43] M. Barrett, "The future of ChatGPT is the future of entertainment," Tech. Rep., Feb. 2023.

[44] S. Sharma, G. Singh, N. Islam, and A. Dhir, "Why do SMEs adopt artificial intelligence-based chatbots?" *IEEE Trans. Eng. Manag.*, 2022.

[45] L. Zhang, Y. Yang, J. Zhou, C. Chen, and L. He, "Retrieval-polished response generation for chatbot," *IEEE Access*, vol. 8, pp. 123882–123890, 2020.

[46] C. Li, X. Zhang, D. Chrysostomou, and H. Yang, "ToD4IR: A humanised task-oriented dialogue system for industrial robots," *IEEE Access*, vol. 10, pp. 91631–91649, 2022.

[47] A. Averza, K. Slhoub, and S. Bhattacharyya, "Evaluating the influence of Twitter bots via agent-based social simulation," *IEEE Access*, vol. 10, pp. 129394–129407, 2022.

[48] Y. Ye, H. You, and J. Du, "Improved trust in human–robot collaboration with ChatGPT," *IEEE Access*, vol. 11, pp. 55748–55754, 2023.

[49] S. Värtinen, P. Hämäläinen, and C. Guckelsberger, "Generating role-playing game quests with GPT language models," *IEEE Trans. Games*, 2022.

[50] G. Cerri, M. Cinalli, F. Michetti, and P. Russo, "Feed forward neural networks for path loss prediction in urban environment," *IEEE Trans. Antennas Propag.*, vol. 52, no. 11, pp. 3137–3139, Nov. 2004.

[51] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang, A. V. Vasilakos, and T. R. Gadekallu, "Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," 2023, *arXiv:2305.10435*.

[52] H. Liu, Y. Cai, Z. Lin, Z. Ou, Y. Huang, and J. Feng, "Variational latent-state GPT for semi-supervised task-oriented dialog systems," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 970–984, 2023.

[53] K. Imoto, T. Nakai, T. Ike, K. Haruki, and Y. Sato, "A CNN-based transfer learning method for defect classification in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 4, pp. 455–459, Nov. 2019.
[54] I. Shalyminov, A. Sordoni, A. Atkinson, and H. Schulz, "GRTr: Generative-retrieval transformers for data-efficient dialogue domain adaptation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2484–2492, 2021.
[55] S. Kim, "A virtual knowledge distillation via conditional GAN," *IEEE Access*, vol. 10, pp. 34766–34778, 2022.
[56] J.-Y. Wu, C. Yu, S.-W. Fu, C.-T. Liu, S.-Y. Chien, and Y. Tsao, "Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1887–1891, Dec. 2019.
[57] T. Yang, F. Ma, X. Li, F. Liu, Y. Zhao, Z. He, and L. Jiang, "DTATrans: Leveraging dynamic token-based quantization with accuracy compensation mechanism for efficient transformer architecture," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 42, no. 2, pp. 509–520, Feb. 2023.
[58] Z. Wang, L. Du, and Y. Li, "Boosting lightweight CNNs through network pruning and knowledge distillation for SAR target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8386–8397, 2021.
[59] Q. Qi, Y. Lu, J. Li, J. Wang, H. Sun, and J. Liao, "Learning low resource consumption CNN through pruning and quantization," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 2, pp. 886–903, Apr. 2022.
[60] M. Wazid, A. K. Das, V. Chamola, and Y. Park, "Uniting cyber security and machine learning: Advantages, challenges and future research," *ICT Exp.*, vol. 8, no. 3, pp. 313–321, Sep. 2022.
[61] E. T. R. Schneider, J. V. A. de Souza, Y. B. Gumiel, C. Moro, and E. C. Paraiso, "A GPT-2 language model for biomedical texts in Portuguese," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2021, pp. 474–479.
[62] M. Abdullah, A. Madain, and Y. Jararweh, "ChatGPT: Fundamentals, applications and social impacts," in *Proc. 9th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Nov. 2022, pp. 1–8.
[63] J. A. Baktash and M. Dawodi, "GPT-4: A review on advancements and opportunities in natural language processing," 2023, *arXiv:2305.03195*.
[64] S. Yang, Q. Li, W. Li, X. Li, and A.-A. Liu, "Dual-level representation enhancement on characteristic and context for image-text retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 8037–8050, Nov. 2022.
[65] Z. Zhang, J. Li, D. G. Stork, E. Mansfield, J. Russell, C. Adams, and J. Z. Wang, "Reducing bias in AI-based analysis of visual artworks," *IEEE BITS Inf. Theory Mag.*, vol. 2, no. 1, pp. 36–48, Oct. 2022.
[66] H. Zhu, P. Tiwari, A. Ghoneim, and M. S. Hossain, "A collaborative AI-enabled pretrained language model for AIoT domain question answering," *IEEE Trans. Ind. Informat.*, vol. 18, no. 5, pp. 3387–3396, May 2022.
[67] J.-H. Syu, J. C.-W. Lin, G. Srivastava, and K. Yu, "A comprehensive survey on artificial intelligence empowered edge computing on consumer electronics," *IEEE Trans. Consum. Electron.*, 2023.
[68] M. Nam, S. Park, and D. S. Kim, "Intrusion detection method using bi-directional GPT for in-vehicle controller area networks," *IEEE Access*, vol. 9, pp. 124931–124944, 2021.
[69] J. An, S. Cho, J. Bang, and M. Kim, "Domain-slot relationship modeling using a pre-trained language encoder for multi-domain dialogue state tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2091–2102, 2022.
[70] L. Medeiros, T. Bosse, and C. Gerritsen, "Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress," *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 3, pp. 343–353, Jun. 2022.
[71] A. Goscinski, E. Bertino, and S. Wang, "Guest editor's introduction: Special section on edge AI as a service," *IEEE Trans. Services Comput.*, vol. 15, no. 2, pp. 588–590, Mar. 2022.
[72] A. Sheth, H. Y. Yip, and S. Shekarpour, "Extending patient-chatbot experience with Internet-of-Things and background knowledge: Case studies with healthcare applications," *IEEE Intell. Syst.*, vol. 34, no. 4, pp. 24–30, Jul. 2019.
[73] L. Hurley, B. S. Kristal, S. Sirimulla, C. Schweikert, and D. F. Hsu, "Multi-layer combinatorial fusion using cognitive diversity," *IEEE Access*, vol. 9, pp. 3919–3935, 2021.
[74] E. S. Maddy and S. A. Boukabara, "MIIDAPS-AI: An explainable machine-learning algorithm for infrared and microwave remote sensing and data assimilation preprocessing—Application to LEO and GEO sensors," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8566–8576, 2021.
[75] C. Park, C. Lee, Y. Yang, and H. Lim, "Ancient Korean neural machine translation," *IEEE Access*, vol. 8, pp. 116617–116625, 2020.
[76] R. Oruche, V. Gundlapalli, A. P. Biswal, P. Calyam, M. L. Alarcon, Y. Zhang, N. R. Bhamidipati, A. Malladi, and H. Regunath, "Evidence-based recommender system for a COVID-19 publication analytics service," *IEEE Access*, vol. 9, pp. 79400–79415, 2021.
[77] H. Qian, X. Li, H. Zhong, Y. Guo, Y. Ma, Y. Zhu, Z. Liu, Z. Dou, and J.-R. Wen, "Pchatbot: A large-scale dataset for personalized chatbot," 2021, *arXiv:2009.13284*.
[78] U. Varolgunes, S. Yao, Y. Ma, and D. Yu, "Embedding imputation with self-supervised graph neural networks," *IEEE Access*, 2023.
[79] M. Jena and S. Dehuri, "An integrated novel framework for coping missing values imputation and classification," *IEEE Access*, vol. 10, pp. 69373–69387, 2022.
[80] B. Islam, M. Iqbal, G. Ubakanma, and S. van der Vliet-Firth, "Skillbot: A conversational chatbot based data mining and sentiment analysis," in *Proc. Human-Centered Cognit. Syst. (HCCS)*, Dec. 2022, pp. 1–10.
[81] A. Miklosik, N. Evans, and A. M. A. Qureshi, "The use of chatbots in digital business transformation: A systematic literature review," *IEEE Access*, vol. 9, pp. 106530–106539, 2021.
[82] J.-C. Gu, Z.-H. Ling, and Q. Liu, "Utterance-to-utterance interactive matching network for multi-turn response selection in retrieval-based chatbots," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 369–379, 2020.
[83] M. Mirzaei, A. Asif, and H. Rivaz, "Combining total variation regularization with window-based time delay estimation in ultrasound elastography," *IEEE Trans. Med. Imag.*, vol. 38, no. 12, pp. 2744–2754, Dec. 2019.
[84] V. Vivek, T. R. Mahesh, C. Saravanan, and K. V. Kumar, "A novel technique for user decision prediction and assistance using machine learning and NLP: A model to transform the e-commerce system," in *Big data management in Sensing: Applications in AI and IoT*. Denmark: River Publishers, 2021, pp. 61–76.
[85] A. Borodin, R. Veynberg, and O. Litvishko, "Methods of text processing when creating chatbots," *Humanitarian Balkan Res.*, vol. 3, no. 5, Aug. 2019.
[86] C. Zhang, L. F. D'Haro, Q. Zhang, T. Friedrichs, and H. Li, "PoE: A panel of experts for generalized automatic dialogue assessment," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1234–1250, 2023.
[87] Y.-T. Lin, A. Papangelis, S. Kim, and D. Hakkani-Tur, "Knowledge-grounded conversational data augmentation with generative conversational networks," 2022, *arXiv:2207.11363*.
[88] Y. Sujana and H.-Y. Kao, "LiDA: Language-independent data augmentation for text classification," *IEEE Access*, vol. 11, pp. 10894–10901, 2023.
[89] Y.-S. Joo, H. Bae, Y.-I. Kim, H.-Y. Cho, and H.-G. Kang, "Effective emotion transplantation in an end-to-end text-to-speech system," *IEEE Access*, vol. 8, pp. 161713–161719, 2020.
[90] C.-S. Jung, Y.-S. Joo, and H.-G. Kang, "Waveform interpolation-based speech analysis/synthesis for HMM-based TTS systems," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 809–812, Dec. 2012.
[91] J. Jeong, S. Cha, J. Choi, S. Yun, T. Moon, and Y. Yoo, "Observations on K-image expansion of image-mixing augmentation," *IEEE Access*, vol. 11, pp. 16631–16643, 2023.
[92] A.-A. Liu, Y. Zhai, N. Xu, W. Nie, W. Li, and Y. Zhang, "Region-aware image captioning via interaction learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3685–3696, Jun. 2022.
[93] S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, "Data augmentation using deep generative models for embedding based speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2598–2609, 2020.
[94] V. Chamola, A. Goyal, P. Sharma, V. Hassija, H. T. T. Binh, and V. Saxena, "Artificial intelligence-assisted blockchain-based framework for smart and secure EMR management," *Neural Comput. Appl.*, vol. 35, no. 31, pp. 22959–22969, Nov. 2023.

[95] K. Leins, M. Cheong, S. Coghlan, S. D'Alfonso, P. Gooding, R. Lederman, and J. Paterson, ''To chat, or bot to chat, just the first question: Potential legal and ethical issues arising from a chatbot case study,'' Tech. Rep., Oct. 2020.

[96] R. B. Saglam, J. R. C. Nurse, and D. Hodges, ''Privacy concerns in chatbot interactions: When to trust and when to worry,'' in *HCI International 2021—Posters*. Cham, Switzerland: Springer, 2021, pp. 391–399.

[97] C. Khatri, B. Hedayatnia, R. Goel, A. Venkatesh, R. Gabriel, and A. Mandal, ''Detecting offensive content in open-domain conversations using two stage semi-supervision,'' 2018, *arXiv:1811.12900*.

[98] N. Naren, V. Chamola, S. Baitragunta, A. Chintanpalli, P. Mishra, S. Yenuganti, and M. Guizani, ''IoMT and DNN-enabled drone-assisted COVID-19 screening and detection framework for rural areas,'' *IEEE Internet Things Mag.*, vol. 4, no. 2, pp. 4–9, Jun. 2021.

[99] K.-M. Kim, W.-J. Ryu, J.-H. Park, and S. Lee, ''MeChat: In-device personal assistant for conversational photo sharing,'' *IEEE Internet Comput.*, vol. 23, no. 2, pp. 23–30, Mar. 2019.

[100] A. Fuad and M. Al-Yahya, ''Recent developments in Arabic conversational AI: A literature review,'' *IEEE Access*, vol. 10, pp. 23842–23859, 2022.

[101] H. Du, S. Teng, H. Chen, J. Ma, X. Wang, C. Gou, B. Li, S. Ma, Q. Miao, X. Na, P. Ye, H. Zhang, G. Luo, and F.-Y. Wang, ''Chat with ChatGPT on intelligent vehicles: An IEEE TIV perspective,'' *IEEE Trans. Intell. Vehicles*, vol. 8, no. 3, pp. 2020–2026, Mar. 2023.

[102] G. Li, L. Liu, C. Zhu, R. Wang, T. Zhao, and S. Shi, ''Detecting source contextual barriers for understanding neural machine translation,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3158–3169, 2021.

[103] K. A. Thakoor, S. C. Koorathota, D. C. Hood, and P. Sajda, ''Robust and interpretable convolutional neural networks to detect glaucoma in optical coherence tomography images,'' *IEEE Trans. Biomed. Eng.*, vol. 68, no. 8, pp. 2456–2466, Aug. 2021.

[104] M. Cao, Y. Fan, Y. Zhang, J. Wang, and Y. Yang, ''VDTR: Video deblurring with transformer,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 160–171, Jan. 2023.

[105] H. S. Nawaz, Z. Shi, Y. Gan, A. Hirpa, J. Dong, and H. Zheng, ''Temporal moment localization via natural language by utilizing video question answers as a special variant and bypassing NLP for corpora,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6174–6185, Sep. 2022.

[106] M. Maitra, D. Saha, P. S. Bhattacharjee, and A. Mukherjee, ''An intelligent paging strategy using rule-based AI technique for locating mobile terminals in cellular wireless networks,'' *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1834–1845, May 2008.

[107] W. Kuang, Y.-L. Chan, S.-H. Tsang, and W.-C. Siu, ''Machine learning-based fast intra mode decision for HEVC screen content coding via decision trees,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 5, pp. 1481–1496, May 2020.

[108] Y. Wang, W. Rong, Y. Ouyang, and Z. Xiong, ''Augmenting dialogue response generation with unstructured textual knowledge,'' *IEEE Access*, vol. 7, pp. 34954–34963, 2019.

[109] F. Wu, M. Du, C. Fan, R. Tang, Y. Yang, A. Mostafavi, and X. Hu, ''Understanding social biases behind location names in contextual word embedding models,'' *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 2, pp. 458–468, 2022.

[110] W. Xu, X. Dong, L. Ma, A. B. J. Teoh, and Z. Lin, ''RawFormer: An efficient vision transformer for low-light RAW image enhancement,'' *IEEE Signal Process. Lett.*, vol. 29, pp. 2677–2681, 2022.

[111] J. Liu, C. Fan, Y. Peng, J. Du, Z. Wang, and C. Chu, ''Emergent leader–follower relationship in networked multiagent systems,'' *Sci. China Inf. Sci.*, vol. 66, no. 12, Dec. 2023, Art. no. 229201.

[112] T.-Y. Chen, Y.-C. Chiu, N. Bi, and R. T. Tsai, ''Multi-modal chatbot in intelligent manufacturing,'' *IEEE Access*, vol. 9, pp. 82118–82129, 2021.

[113] L. Nie, F. Jiao, W. Wang, Y. Wang, and Q. Tian, ''Conversational image search,'' *IEEE Trans. Image Process.*, vol. 30, pp. 7732–7743, 2021.

[114] D. Streeb, Y. Metz, U. Schlegel, B. Schneider, M. El-Assady, H. Neth, M. Chen, and D. A. Keim, ''Task-based visual interactive modeling: Decision trees and rule-based classifiers,'' *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 9, pp. 3307–3323, Sep. 2022.

[115] J. M. Mendel and P. P. Bonissone, ''Critical thinking about explainable AI (XAI) for rule-based fuzzy systems,'' *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3579–3593, Dec. 2021.

[116] S. R. Gaddam, V. V. Phoha, and K. S. Balagani, ''K-means+ID3: A novel method for supervised anomaly detection by cascading k-means clustering and ID3 decision tree learning methods,'' *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 345–354, Mar. 2007.

[117] J.-S. Lee, ''AUC4.5: AUC-based C4.5 decision tree algorithm for imbalanced data classification,'' *IEEE Access*, vol. 7, pp. 106034–106042, 2019.

[118] P. Chhikara, R. Tekchandani, N. Kumar, V. Chamola, and M. Guizani, ''DCNN-GA: A deep neural net architecture for navigation of UAV in indoor environment,'' *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4448–4460, Mar. 2021.

[119] B. Cheng, M. Wang, S. Zhao, Z. Zhai, D. Zhu, and J. Chen, ''Situation-aware dynamic service coordination in an IoT environment,'' *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2082–2095, Aug. 2017.

[120] C. Guo, W. Ai, S. Hu, X. Du, and N. Chen, ''Sea surface wind direction retrieval based on convolution neural network and wavelet analysis,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3868–3876, 2022.

[121] Y. R. Choi and R. M. Kil, ''Face video retrieval based on the deep CNN with RBF loss,'' *IEEE Trans. Image Process.*, vol. 30, pp. 1015–1029, 2021.

[122] Y. Lai, X. Dong, Z. Jin, M. Tistarelli, W.-S. Yap, and B.-M. Goi, ''Breaking free from entropy's shackles: Cosine distance-sensitive error correction for reliable biometric cryptography,'' *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3101–3115, 2023.

[123] H. Yan and Y. Tang, ''Collaborative filtering based on Gaussian mixture model and improved Jaccard similarity,'' *IEEE Access*, vol. 7, pp. 118690–118701, 2019.

[124] K. Palasundram, N. M. Sharef, K. A. Kasmiran, and A. Azman, ''Enhancements to the sequence-to-sequence-based natural answer generation models,'' *IEEE Access*, vol. 8, pp. 45738–45752, 2020.

[125] K. Palasundram, N. M. Sharef, K. A. Kasmiran, and A. Azman, ''SEQ2SEQ++: A multitasking-based Seq2Seq model to generate meaningful and relevant answers,'' *IEEE Access*, vol. 9, pp. 164949–164975, 2021.

[126] X. Li, F. Zhou, C. Xu, J. Ji, and G. Yang, ''SEA: Sentence encoder assembly for video retrieval by textual queries,'' *IEEE Trans. Multimedia*, vol. 23, pp. 4351–4362, 2021.

[127] B. Zhang, D. Xiong, J. Su, and H. Duan, ''A context-aware recurrent encoder for neural machine translation,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2424–2432, Dec. 2017.

[128] S. S. Kusumawardani and S. A. I. Alfarozi, ''Transformer encoder model for sequential prediction of student performance based on their log activities,'' *IEEE Access*, vol. 11, pp. 18960–18971, 2023.

[129] A. Sharif, G. Zhai, X. Min, J. Jia, and K. Munir, ''Enhancing decoding rate of barcode decoders in complex scenes for IoT systems,'' *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17495–17507, Dec. 2021.

[130] T. Jalaja, Dr. T. Adilakshmi, M. S. S. Chandra, M. I. Mirza, and M. Kumar, ''A behavioral chatbot using encoder–decoder architecture: Humanizing conversations,'' in *Proc. 2nd Int. Conf. Interdiscipl. Cyber Phys. Syst. (ICPS)*, May 2022, pp. 51–54.

[131] A. Wu, Y. Han, Z. Zhao, and Y. Yang, ''Hierarchical memory decoder for visual narrating,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2438–2449, Jun. 2021.

[132] W. Wang, A. Wang, Q. Ai, C. Liu, and J. Liu, ''AAGAN: Enhanced single image dehazing with attention-to-attention generative adversarial network,'' *IEEE Access*, vol. 7, pp. 173485–173498, 2019.

[133] G. Mao, J. Su, S. Yu, and D. Luo, ''Multi-turn response selection for chatbots with hierarchical aggregation network of multi-representation,'' *IEEE Access*, vol. 7, pp. 111736–111745, 2019.

[134] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, ''STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG,'' *IEEE Trans. Biomed. Eng.*, vol. 69, no. 7, pp. 2233–2242, Jul. 2022.

[135] Y. Djenouri, A. Belhadi, A. Yazidi, G. Srivastava, and J. C. Lin, ''Artificial intelligence of medical things for disease detection using ensemble deep learning and attention mechanism,'' *Expert Syst.*, Jun. 2022, Art. no. e13093.

[136] M. Murtaza, Y. Ahmed, J. A. Shamsi, F. Sherwani, and M. Usman, ''AI-based personalized e-learning systems: Issues, challenges, and solutions,'' *IEEE Access*, vol. 10, pp. 81323–81342, 2022.

[137] J. M. Gayed, M. K. J. Carlon, A. M. Oriola, and J. S. Cross, ''Exploring an AI-based writing assistant's impact on English language learners,'' *Comput. Educ., Artif. Intell.*, vol. 3, 2022, Art. no. 100055.

[138] T. Kalampokas, K. Tziridis, N. Kalampokas, A. Nikolaou, E. Vrochidou, and G. A. Papakostas, ''A holistic approach on airfare price prediction using machine learning techniques,'' *IEEE Access*, vol. 11, pp. 46627–46643, 2023.

[139] J. Wang, Y. Yang, Q. Liu, Z. Fang, S. Sun, and Y. Xu, ''An empirical study of user engagement in influencer marketing on Weibo and WeChat,'' *IEEE Trans. Computat. Social Syst.*, pp. 1–13, 2022.

[140] C. Li, D. Chrysostomou, X. Zhang, and H. Yang, ''IRWoZ: Constructing an industrial robot Wizard-of-OZ dialoguing dataset,'' *IEEE Access*, vol. 11, pp. 28236–28251, 2023.

[141] Z. Hu, Z. Cao, H. P. Chan, J. Liu, X. Xiao, J. Su, and H. Wu, ''Controllable dialogue generation with disentangled multi-grained style specification and attribute consistency reward,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 188–199, 2023.

[142] M. Rizinski, H. Peshov, K. Mishev, L. T. Chitkushev, I. Vodenska, and D. Trajanov, ''Ethically responsible machine learning in fintech,'' *IEEE Access*, vol. 10, pp. 97531–97554, 2022.

[143] T. Alladi, B. Gera, A. Agrawal, V. Chamola, and F. R. Yu, ''DeepADV: A deep neural network framework for anomaly detection in VANETs,'' *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 12013–12023, Nov. 2021.

[144] I. Hussain, R. Ahmad, S. Muhammad, K. Ullah, H. Shah, and A. Namoun, ''PHTI: Pashto handwritten text imagebase for deep learning applications,'' *IEEE Access*, vol. 10, pp. 113149–113157, 2022.

[145] S. Singh, N. Wen, Y. Hou, P. Alipoormolabashi, T.-L. Wu, X. Ma, and N. Peng, ''COM2SENSE: A commonsense reasoning benchmark with complementary sentences,'' 2021, *arXiv:2106.00969*.

[146] Q. Zhou, Z. Feng, Q. Gu, J. Pang, G. Cheng, X. Lu, J. Shi, and L. Ma, ''Context-aware mixup for domain adaptive semantic segmentation,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 804–817, Feb. 2023.

[147] A. Yadav, A. Patel, and M. Shah, ''A comprehensive review on resolving ambiguities in natural language processing,'' *AI Open*, vol. 2, pp. 85–92, Jan. 2021.

[148] X. Liao, Y. Huang, Y. Wei, C. Zhang, F. Wang, and Y. Wang, ''Efficient estimate of sentence's representation based on the difference semantics model,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3384–3399, 2021.

[149] Y. Zhao, R. Cao, J. Bai, W. Ma, and H. Shinnou, ''Determining the logical relation between two sentences by using the masked language model of BERT,'' in *Proc. Int. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Dec. 2020, pp. 228–231.

[150] B. Bhowmik, P. Hazarika, P. Kale, and S. Jain, ''AI technology for NoC performance evaluation,'' *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 12, pp. 3483–3487, Dec. 2021.

[151] B. Kim, J. Seo, and M.-W. Koo, ''Randomly wired network based on RoBERTa and dialog history attention for response selection,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2437–2442, 2021.

[152] A. Abdellatif, K. Badran, D. E. Costa, and E. Shihab, ''A comparison of natural language understanding platforms for chatbots in software engineering,'' *IEEE Trans. Softw. Eng.*, vol. 48, no. 8, pp. 3087–3102, Aug. 2022.

[153] W. Cai, Y. Jin, and L. Chen, ''Task-oriented user evaluation on critiquing-based recommendation chatbots,'' *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 3, pp. 354–366, Jun. 2022.

[154] H. Tanaka, H. Iwasaka, Y. Matsuda, K. Okazaki, and S. Nakamura, ''Analyzing self-efficacy and summary feedback in automated social skills training,'' *IEEE Open J. Eng. Med. Biol.*, vol. 2, pp. 65–70, 2021.

[155] G. A. Santos, G. G. de Andrade, G. R. S. Silva, F. C. M. Duarte, J. P. J. da Costa, and R. T. de Sousa, ''A conversation-driven approach for chatbot management,'' *IEEE Access*, vol. 10, pp. 8474–8486, 2022.

[156] A. Mazzei, L. Anselma, M. Sanguinetti, A. Rapp, D. Mana, M. M. Hossain, V. Patti, R. Simeoni, and L. Longo, ''Anticipating user intentions in customer care dialogue systems,'' *IEEE Trans. Human-Mach. Syst.*, vol. 52, no. 5, pp. 973–983, Oct. 2022.

[157] J. Choe, S. J. Oh, S. Chun, S. Lee, Z. Akata, and H. Shim, ''Evaluation for weakly supervised object localization: Protocol, metrics, and datasets,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1732–1748, Feb. 2023.

[158] Y. Zhang, W. Zhou, M. Wang, Q. Tian, and H. Li, ''Deep relation embedding for cross-modal retrieval,'' *IEEE Trans. Image Process.*, vol. 30, pp. 617–627, 2021.

[159] A.-H. Al-Ajmi and N. Al-Twairesh, ''Building an Arabic flight booking dialogue system using a hybrid rule-based and data driven approach,'' *IEEE Access*, vol. 9, pp. 7043–7053, 2021.

[160] S. Ni and H.-Y. Kao, ''Masked Siamese prompt tuning for few-shot natural language understanding,'' *IEEE Trans. Artif. Intell.*, 2023.

[161] J.-H. Lee, E. H. Wu, Y.-Y. Ou, Y.-C. Lee, C.-H. Lee, and C.-R. Chung, ''Anti-drugs chatbot: Chinese BERT-based cognitive intent analysis,'' *IEEE Trans. Computat. Social Syst.*, 2023.

[162] L. Zhang and Y. Hu, ''A fine-tuning approach research of pre-trained model with two stage,'' in *Proc. IEEE Int. Conf. Power Electron., Comput. Appl. (ICPECA)*, Jan. 2021, pp. 905–908.

[163] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, ''Chatting and cheating: Ensuring academic integrity in the era of ChatGPT,'' *Innov. Educ. Teaching Int.*, pp. 1–12, Mar. 2023, doi: 10.1080/14703297.2023.2190148.

[164] T. Susnjak, ''ChatGPT: The end of online exam integrity?'' 2022, *arXiv:2212.09292*.

[165] B. Zierock, ''ChatGPT and content creation automation,'' Tech. Rep., Jan. 2023.

[166] F. Meng, X. Xiao, and J. Wang, ''Rating the crisis of online public opinion using a multi-level index system,'' *Int. Arab J. Inf. Technol.*, vol. 19, no. 4, pp. 597–608, 2022.

[167] M. Haman and M. Školník, ''Using ChatGPT to conduct a literature review,'' *Accountability Res.*, pp. 1–3, Mar. 2023.

[168] E. Cambiaso and L. Caviglione, ''Scamming the scammers: Using ChatGPT to reply mails for wasting time and resources,'' 2023, *arXiv:2303.13521*.

[169] M. Thisarani and S. Fernando, ''Artificial intelligence for futuristic banking,'' in *Proc. IEEE Int. Conf. Eng., Technol. Innov. (ICE/ITMC)*, Jun. 2021, pp. 1–13.

[170] M. Abbas, ''Uses and misuses of ChatGPT by academic community: An overview and guidelines,'' *SSRN Electron. J.*, pp. 1–13, Apr. 2023.

[171] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and Y. Liu, ''Jailbreaking chatgpt via prompt engineering: An empirical study,'' 2023.

[172] T. Alladi, V. Chamola, N. Sahu, V. Venkatesh, A. Goyal, and M. Guizani, ''A comprehensive survey on the applications of blockchain for securing vehicular networks,'' *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1212–1239, 1st Quart., 2022.

[173] T. Alladi, A. Agrawal, B. Gera, V. Chamola, B. Sikdar, and M. Guizani, ''Deep neural networks for securing IoT enabled vehicular ad-hoc networks,'' in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

[174] A. Yazdinejad, A. Dehghantanha, and G. Srivastava, ''AP2FL: Auditable privacy-preserving federated learning framework for electronics in healthcare,'' *IEEE Trans. Consum. Electron.*, 2023.

[175] H. Zhang, Y. Mi, Y. Fu, X. Liu, Y. Zhang, J. Wang, and J. Tan, ''Security defense decision method based on potential differential game for complex networks,'' *Comput. Secur.*, vol. 129, Jun. 2023, Art. no. 103187.

[176] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, ''The emerging threat of AI-driven cyber attacks: A review,'' *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2037254.

[177] P. Chen, H. Liu, R. Xin, T. Carval, J. Zhao, Y. Xia, and Z. Zhao, ''Effectively detecting operational anomalies in large-scale IoT data infrastructures by using a GAN-based predictive model,'' *Comput. J.*, vol. 65, no. 11, pp. 2909–2925, Nov. 2022.

[178] H. Grover, T. Alladi, V. Chamola, D. Singh, and K. R. Choo, ''Edge computing and deep learning enabled secure multitier network for Internet of Vehicles,'' *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14787–14796, Oct. 2021.

[179] B.-H. Juang and S. Furui, ''Automatic recognition and understanding of spoken language—A first step toward natural human–machine communication,'' *Proc. IEEE*, vol. 88, no. 8, pp. 1142–1165, Aug. 2000.

[180] S. AlZu'bi, A. Mughaid, F. Quiam, and S. Hendawi, ''Exploring the capabilities and limitations of ChatGPT and alternative big language models,'' *Artif. Intell. Appl.*, pp. 1–5, Apr. 2023.

[181] A. Følstad, P. Brandtzaeg, T. Feltwell, E. Law, M. Tscheligi, and E. Luger, ''SIG: Chatbots for social good,'' in *Proc. Extended Abstr. CHI Conf.*, Apr. 2018, pp. 1–4.

[182] N. Chowdhury and S. Rahman, ''A brief review of ChatGPT: Limitations, challenges and ethical-social implications,'' Tech. Rep., Feb. 2023.

[183] B. Xing and I. W. Tsang, ''Understand me, if you refer to aspect knowledge: Knowledge-aware gated recurrent memory network,'' *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1092–1102, Oct. 2022.

[184] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A SWOT analysis of ChatGPT: Implications for educational practice and research," *Innov. Educ. Teaching Int.*, pp. 1–15, Mar. 2023, doi: 10.1080/14703297.2023.2195846.

[185] I. AlAgha, "Multihop question answering by using sequential path expansion with backtracking," *IEEE Access*, vol. 10, pp. 76842–76854, 2022.

[186] S. Fergus, M. Botha, and M. Ostovar, "Evaluating academic answers generated using ChatGPT," *J. Chem. Educ.*, vol. 100, no. 4, pp. 1672–1675, Apr. 2023, doi: 10.1021/acs.jchemed.3c00087.

[187] H. Ibrahim, R. Asim, F. Zaffar, T. Rahwan, and Y. Zaki, "Rethinking homework in the age of artificial intelligence," *IEEE Intell. Syst.*, vol. 38, no. 2, pp. 24–27, Mar. 2023.

[188] K. Kaur, T. Thanuja, O. Dahiya, P. T. Sai, H. Kaur, and J. Singh, "Design and development of a ticket booking system using smart bot," in *Proc. 10th Int. Conf. Rel., INFOCOM Technol. Optim., Trends Future Directions (ICRITO)*, Oct. 2022, pp. 1–6.

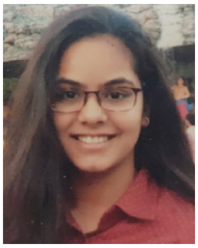
[189] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, "Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models," 2023, *arXiv:2304.01852*.

[190] E. Mangina, *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report—Social and Multi-User Spaces in VR: Trolling, Harassment, and Online Safety*, IEEE Standard, 2021, pp. 1–17.

[191] B. D. Lund, T. Wang, N. R. Mannuru, B. Nie, S. Shimray, and Z. Wang, "ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing," *J. Assoc. Inf. Sci. Technol.*, pp. 1–23, Mar. 2023, doi: 10.1002/asi.24750.

VIKAS HASSIJA received the B.Tech. degree from M.D.U. University, Rohtak, India, in 2010, the M.S. degree in telecommunications and software engineering from the Birla Institute of Technology and Science (BITS), Pilani, India, in 2014, and the Ph.D. degree in IoT security and blockchain from the Jaypee Institute of Information and technology (JIIT), Noida. He was an Assistant Professor with JIIT for four years. He is currently a Postdoctoral Researcher with the National University of Singapore, Singapore. He has eight years of industry experience and has worked with various telecommunication companies like Tech Mahindra and Accenture. His research interests include the IoT security, network security, blockchain, and distributed computing.

ARJAB CHAKRABARTI is currently pursuing the B.Tech. degree with the Kalinga Institute of Industrial Technology (KIIT), Bhubaneshwar. He has completed a few projects in the field of machine learning and web designing. He is also doing his research internship with the Birla Institute of Technology and Science (BITS), Pilani, under the supervision of Dr. Vikas Hassija. His research interests include machine learning, reinforcement learning, quantum computing, and deep learning.



ANUSHKA SINGH is currently pursuing the B.Tech. degree with the Kalinga Institute of Industrial Technology (KIIT), Bhubaneshwar. She is also doing her research internship with the Birla Institute of Technology and Science (BITS), Pilani, under the supervision of Dr. Vikas Hassija. Her research interests include machine learning, reinforcement learning, quantum computing, and deep learning.

VINAY CHAMOLA (Senior Member, IEEE) received the B.E. degree in electrical and electronics engineering and the master's degree in communication engineering from the Birla Institute of Technology and Science, Pilani, India, in 2010 and 2013, respectively, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2016. In 2015, he was a Visiting Researcher with the Autonomous Networks Research Group (ANRG), University of Southern California, Los Angeles, CA, USA. He was a Postdoctoral Research Fellow with the National University of Singapore, Singapore. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, BITS Pilani, where he heads the Internet of Things Research Group/Laboratory. He is the Co-Founder and the President of a healthcare startup Medsupervision Pvt. Ltd. His research interests include the IoT security, blockchain, UAVs, VANETs, 5G, and healthcare. He is a fellow of IET. He serves as an Area Editor for the *Ad Hoc Networks* journal (Elsevier) and *IEEE Internet of Things Magazine*. He also serves as an Associate Editor for IEEE Transactions on Intelligent Transportation Systems, IEEE Networking Letters, *IEEE Consumer Electronics Magazine*, *IET Quantum Communications*, *IET Networks*, and several other journals. He serves as the Co-Chair for various reputed workshops like IEEE Globecom Workshop 2021, IEEE INFOCOM 2022 Workshop, IEEE ANTS 2021, and IEEE ICIAfS 2021, to name a few. He is listed in the world's top 2% scientists identified by Stanford University.

BIPLAB SIKDAR (Senior Member, IEEE) received the B.Tech. degree in electronics and communication engineering from North Eastern Hill University, Shillong, India, in 1996, the M.Tech. degree in electrical engineering from the Indian Institute of Technology Kanpur, Kanpur, India, in 1998, and the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2001. He was a Faculty Member of the Rensselaer Polytechnic Institute, from 2001 to 2013. He was also an Assistant Professor and an Associate Professor. He is currently a Professor and the Head of the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His current research interests include wireless networks and security for Internet of Things and cyber physical systems. He has served as an Associate Editor for IEEE Transactions on Communications, IEEE Transactions on Mobile Computing, IEEE Internet of Things Journal, and IEEE Open Journal of Vehicular Technology.