Demystifying ChatGPT: An In-depth Survey of OpenAI's Robust Large Language Models

Pronaya Bhattacharya[1] · Vivek Kumar Prasad[2] · Ashwin Verma[2] · Deepak Gupta[3] · Assadaporn Sapsomboon[4] · Wattana Viriyasitavat[4] · Gaurav Dhiman[5,6,7,8,9,10]

Abstract

Recent advancements in natural language processing (NLP) have catalyzed the development of models capable of generating coherent and contextually relevant responses. Such models are applied across a diverse array of applications, including but not limited to chatbots, expert systems, question-and-answer robots, and language translation systems. Large Language Models (LLMs), exemplified by OpenAI's Generative Pretrained Transformer (GPT), have significantly transformed the NLP landscape. They have introduced unparalleled abilities in generating text that is not only contextually appropriate but also semantically rich. This evolution underscores a pivotal shift towards more sophisticated and intuitive language understanding and generation capabilities within the field. Models based on GPT are developed through extensive training on vast datasets, enabling them to grasp patterns akin to human writing styles and deliver insightful responses to intricate questions. These models excel in condensing text, extending incomplete passages, crafting imaginative narratives, and emulating conversational exchanges. However, GPT LLMs are not without their challenges, including ethical dilemmas and the propensity for disseminating misinformation. Additionally, the deployment of these models at a practical scale necessitates a substantial investment in training and computational resources, leading to concerns regarding their sustainability. ChatGPT, a variant rooted in transformer-based architectures, leverages a self-attention mechanism for data sequences and a reinforcement learning-based human feedback (RLHF) system. This enables the model to grasp long-range dependencies, facilitating the generation of contextually appropriate outputs. Despite ChatGPT marking a significant leap forward in NLP technology, there remains a lack of comprehensive discourse on its architecture, efficacy, and inherent constraints. Therefore, this survey aims to elucidate the ChatGPT model, offering an in-depth exploration of its foundational structure and operational efficacy. We meticulously examine Chat-GPT's architecture and training methodology, alongside a critical analysis of its capabilities in language generation. Our investigation reveals ChatGPT's remarkable aptitude for producing text indistinguishable from human writing, whilst also acknowledging its limitations and susceptibilities to bias. This analysis is intended to provide a clearer understanding of ChatGPT, fostering a nuanced appreciation of its contributions and challenges within the broader NLP field. We also explore the ethical and societal implications of this technology, and discuss the future of NLP and AI. Our study provides valuable insights into the inner workings of ChatGPT, and helps to shed light on the potential of LLMs for shaping the future of technology and society. The approach used as Eco-GPT, with a three-level cascade (GPT-J, J1-G, GPT-4), achieves 73% and 60% cost savings in CaseHold and CoQA datasets, outperforming GPT-4.

1 Introduction

The advent of Large Language Models (LLMs) has marked a paradigm shift in the field of Natural Language Processing (NLP) [1–6], transforming how we interact with and understand textual data [7]. Among the pioneers in this domain, OpenAI's Generative Pre-Trained Transformers (GPT) have emerged as a prominent force, reshaping the landscape of NLP [8]. Comprising multiple transformer layers, GPT models employ self-attention mechanisms that allow for the intricate comprehension of language patterns within extensive datasets [9]. This design allows the models to learn semantic meaning from data, thereby achieving remarkable progress in textual understanding. Recent research has delved into the varied applications of GPT technology across multiple sectors, including finance, healthcare, and law. This exploration underscores the wide-ranging applicability and potential of GPT models, highlighting their versatility in addressing

---

Extended author information available on the last page of the article

distinct industry-specific challenges and requirements [10, 11]. Table 1 presents the list of acronyms used throughout the survey.

To illustrate the market penetration of ChatGPT and its integration with users, we present statistical details. Figure 1 demonstrates the efficacy of the ChatGPT website [12] over the past seven months. This observation is substantiated by the graphical representation, which indicates a notable upswing in the utility of ChatGPT. Specifically, in November 2022, the platform garnered 152.7 million views [13–20]. The number of views escalated to 266 million in December 2022, and then saw a more significant rise to 616 million in January 2023. This upward trend persisted, reaching 1 billion views in February, before surging to 1.6 billion in March 2023. Remarkably, by April 2023, user engagement reached its zenith at 1.8 billion, continuing its upward trajectory into May with a total of 1.9 billion views. These
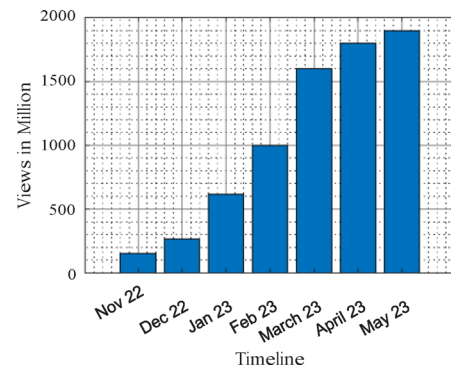

Fig. 1 Monthly view of the ChatGPT website

statistics illustrate the increasing integration of ChatGPT into daily activities, highlighting the growing necessity to enhance the performance of GPT LLMs. Furthermore, additional data reveals ChatGPT's widespread impact across various industries, underscoring its broad applicability and influence.

Figure 2 illustrates the rapid growth of the chatbot sector [12], reaching 100 million users in only two months. This remarkable achievement, when compared to other social media and entertainment platforms, showcases a stark contrast in user base growth. For example, TikTok reached the same number of users within 9 months, whereas platforms such as YouTube, Instagram, Facebook, and Twitter took 18, 30, 54, and 60 months, respectively, to achieve similar milestones. In the music industry, Spotify required a considerably longer period, taking 132 months to reach this mark. Meanwhile, Netflix, a giant in the entertainment sector, achieved the 100 million user milestone in 216 months, further highlighting the unprecedented rapid adoption rate of the platform in question. [21–28].

Since their introduction, GPT models have undergone swift and significant development, beginning with the launch

Table 1 Acronyms and their descriptions

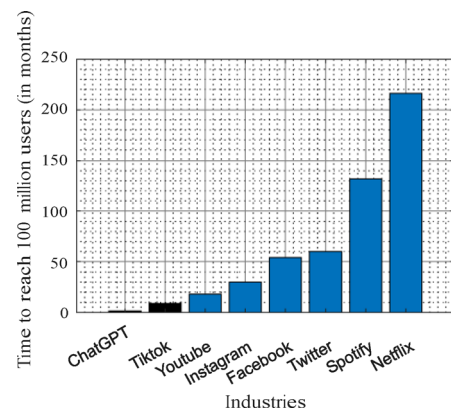| Acronyms	Description
LLMs	Large Language Models
NLP	Natural Language Processing
GPT	Generative Pretrained Transformers
ML	Machine Learning
RLHF	Reinforcement Learning Based Human Feedback
GAI	Generative Artificial Intelligence
SC	Semantic Communication
NLG	Natural Language Generation
PTMs	Pretrained Models
GAI	Generative Artificial Intelligence
ML	Machine Learning
ABL	Abductive Learning
FOL	First Order Logic
FL	Federated Learning
NMT	Neural Machine Translation
ASR	Automatic Speech Recognition
TTS	Text-To-Speech
MTL	Multi-Task Learning
CV	Computer Vision
FFN	Feed-Forward Neural Network
ReLU	Rectified Linear Unit
API	Application Programming Interfaces
IoT	Internet-of-Things
DL	Deep Learning
HCI	Human–Computer Interaction
GPUs	Graphics Processing Unit
TPUs	Tensor Processing Units
RL	Reinforcement Learning

Fig. 2 Impact of ChatGPT in Diverse Industries

of GPT-1, which established the foundation for a groundbreaking series of models in natural language processing (NLP). This evolution has been instrumental in transforming the landscape of NLP, setting new benchmarks for language understanding and generation. [29]. Following the initial model, subsequent iterations such as GPT-2 and GPT-3 have significantly broadened the scope and enhanced the capabilities of these systems, pushing the frontiers of Machine Learning (ML) and language comprehension to new heights. These advancements have paved the way for more sophisticated applications in various domains, showcasing the progressive potential of GPT models in understanding and generating human-like text. [9]. With the recent unveiling of GPT-4 and the anticipation of future models, the GPT series continues to push the frontiers of NLP, offering unprecedented levels of accuracy and efficiency in language tasks [30]. The evolution from GPT-1 to GPT-4 marks an ongoing endeavor to refine the models' complexity, responsiveness, and contextual awareness. This journey is pivotal in shaping the future landscape of interactions between machines and humans, demonstrating a commitment to advancing the capabilities of these models to understand and engage with human language more effectively. [31].

Table 2 provides a comprehensive comparison of different generations of GPT models, encompassing GPT (or GPT-1) to GPT-4. The text outlines the evolution of the models in terms of their parameters, underscoring the escalation in complexity and improvements in self-attention mechanisms. Additionally, it traces the incorporation of Reinforcement Learning based Human Feedback (RLHF), illustrating a shift towards more advanced and nuanced training methodologies. This progression reflects a concerted effort to refine the models' ability to process and understand language through increasingly sophisticated techniques. [32–40]. Furthermore, the table illustrates the diversification in applications, demonstrating the growing versatility and adaptability of GPT models in various domains. (Fig. 3).

The core of GPT models' success lies in the self-attention mechanism, an innovative technique that has revolutionized ML models ability to recognize dependencies across sequences in data [41]. The self-attention mechanism allows each part of an input sequence to focus on different parts of itself, enabling the model to create contextual representations. Alongside self-attention, RLHF adds another dimension to the model's training, aligning the generated content with human-like reasoning and judgment [42]. By fine-tuning models using RLHF, it is possible to guide the model's behavior based on human-generated feedback, optimizing performance in specialized domains [43]. This dual application of self-attention and RLHF has contributed to the efficiency of various applications such as text summarization, creative writing, and question-answering systems [44]. For instance, GPT-3 ability to perform legal document analysis has been benchmarked against human experts, revealing astonishing parallels in terms of accuracy and insight [11, 41]. Moreover, the use of self-attention and RLHF has been pivotal in addressing the long-standing challenge of long-range dependencies in sequence data, leading to breakthroughs in machine translation and other complex tasks [45].

The training process of these models is both complex and resource-intensive, leveraging vast amounts of data and specialized computational hardware [46]. By employing a multi-stage approach that combines pre-training on broad data and fine-tuning on specific tasks, GPT models are capable of delivering customized performance across a wide range of applications [47]. This training architecture underscores the adaptability and specialization of GPT models, whilst also illuminating the considerable infrastructure and computational requirements necessary for deployment [48]. The synergistic application of self-attention and RLHF, combined with an intricate training process, encapsulates the scientific innovation and practical challenges inherent in the development and implementation of these advanced language models [49–56].

Despite these remarkable advancements, the implementation of GPT models is not without challenges and concerns. The computational cost associated with training these models is substantial, often requiring highly specialized hardware and vast amounts of energy [44, 57]. This raises significant questions about their sustainability and environmental impact, with recent studies highlighting the

Table 2 Comparison of Different Generations of GPT Models

| Model	Parameters	Self-Attention Features	RLHF	Major Applications	Training Data Size	Inference Speed	Year
GPT	117 million	Basic mechanism	No	Text Completion, Translation	40 GB	Fast	2018
GPT-2	1.5 billion	Scaled attention, Layer normalization	No	Creative Writing, Summarization	570 GB	Moderate	2019
GPT-3	175 Billion	Adaptive attention, Sparse attention	Partial	Question Answering, Legal Analysis	45 TB	Slow	2020
GPT-4	1.75 trillion	Enhanced features, Multi-modal capabilities	Yes	Diverse Tasks, Customization, Real-time Interaction	240 TB	Moderate	2022
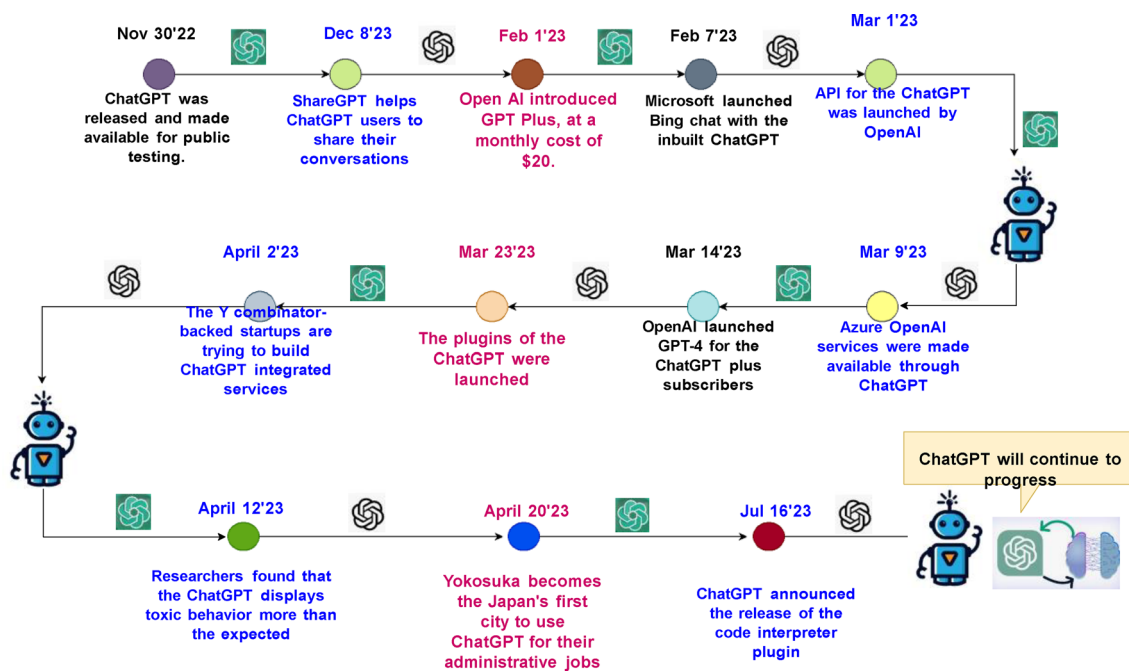
Fig. 3 The evolution timeline of ChatGPT

considerable carbon footprint associated with training large-scale models [58, 59]. Ethically, the deployment of GPT models has stirred debate around issues such as bias, misinformation, and potential misuse. Research has indicated that these models can inadvertently propagate stereotypes or misrepresentations found within their training data, posing serious ethical considerations [60, 61].

Additionally, the complex architecture and extensive resource requirements of GPT models may limit accessibility for researchers and organizations with constrained resources, raising questions about inclusivity and equitable access to this cutting-edge technology [62, 63]. Notably, the applications of GPT models extend into areas like real-time language translation, medical diagnosis support, and personalized education, underscoring the potential societal impact and the necessity for responsible development and deployment [64, 65]. The intersection of ethical considerations, practical constraints, and far-reaching potential paints a complex picture of the current state of LLMs, necessitating ongoing research, collaboration, and critical evaluation.

Therefore, this article aims to demystify the ChatGPT model, bridging the knowledge gap and contributing a holistic understanding of its architecture, performance, and societal implications. In undertaking this endeavor, the analysis leverages recent research, empirical evidence, and a thoughtful scrutiny of the ethical considerations involved. The objective is to furnish scholars, practitioners, and policymakers with a thorough and groundbreaking examination of ChatGPT, offering insights that span theoretical underpinnings, practical applications, and ethical implications, elucidating its role in shaping the future of technology and society. Through a detailed examination of ChatGPT's complexities, this study not only enriches academic dialogue but also acts as a roadmap for future innovations and ethical deployment. Combining theoretical insights with practical applications, it contributes to the broader endeavor of leveraging AI's potential responsibly, with an acute consciousness of its intricacies and obligations. The following sections will investigate ChatGPT's architecture, its diverse applications, ethical implications, and forward-looking views, thereby setting the stage for a well-informed and thorough grasp of this crucial technology.

1.1 Existing Surveys

This section presents the discussion on existing surveys conducted in different research areas by the researchers. Mondal et al. [66] surveyed the problems that Generative Artificial Intelligence (GAI) and its effects on the economy and society. It exhorts companies to modify their tactics in order to combine physical and virtual meetings through GAI integration. They also provide business managers with a useful foundation for creating successful plans. By identifying the potential for businesses to create value propositions and customer experiences through digital communication and information technology, this also examines the GAIs initial deployments in diverse industries aims to inspire future customer solutions.

Another study by [67], examined the three main areas of the ChatGPT, namely, the ChatGPT responses to questions pertaining to science education; potential Chat-GPT uses in science pedagogy; and the study of ChatGPT as a research tool and its utility. This research, which used a self-study technique, produced outstanding results, with ChatGPT frequently aligning its outputs with the investigation's main themes. However, this is also important to recognize that ChatGPT runs the risk of taking the position of the supreme epistemic authority. The analysis presents insights that, while unique, lack substantial empirical evidence or the necessary qualifications, a point that is elaborated upon within the survey. From an application-oriented perspective, the use of ChatGPT across various sectors such as healthcare, clinical practice, education, and manufacturing industries is examined, highlighting its versatility and potential impact in these fields [68–74]. The changes to transformer-based LLMs is presented in the article.

In [75], the authors examined the current trends of GAI models, and associated integrations with the metaverse for precision medicine. The salient advantages and inherent drawbacks are highlighted. GAI is a dynamic branch of AI that can create different content types on its own, such as text, photos, audio, and video. GAI presents fresh opportunities for content production within the metaverse [76], successfully overcoming technological gaps in this rapidly developing digital space. Utilizing ChatGPT can markedly enhance the generation and presentation of virtual objects within the metaverse, potentially steering Industry 5.0 towards a new era of hyper-customization and individualized user experiences. This advancement underscores the technology's capacity to meet specific user demands through personalized and dynamic virtual environments.

Lv et al. [77] provides an in-depth analysis of how Generative AI (GAI) technologies are employed to create innovative and captivating content in digital and virtual settings. It outlines the diverse applications of GAI, such as converting text into images, audio, and scientific materials, delivering a thorough review of its capabilities in augmenting creative outputs across various formats. In the study [78], the authors showcased the utilization of Generative AI (GAI) in text-to-code tasks, revealing its capacity for economic optimization and the potential to enhance both creative and non-creative endeavors. However, their current framework suffers from notable limitations, including prolonged training periods due to the unavailability of datasets for certain models like text-to-text and text-to-audio. Additionally, these models demand extensive datasets and substantial computational resources, resulting in significant time and expenses for training and execution [79–85].

Authors in [86] conducted a survey on the application of ChatGPT in education. Initially, the study explores the potential of GPT in academic tasks. Subsequently, it delves into the limitations of GPT, particularly in terms of providing erroneous information and amplifying biases. Furthermore, the survey emphasizes the importance of addressing security concerns during data training and prioritizing privacy considerations. To enhance students' learning experiences, collaborative efforts among policymakers, researchers, educators, and technology experts are deemed essential. The survey encourages discussions on the safe and beneficial integration of emerging Generative AI (GAI) tools in educational settings.

Xia et al. [87] delved into the potential of Generative AI (GAI) technologies in advancing communication networks, specifically within the framework of Web 3.0 cognitive architectures. Their research emphasized the importance of Semantic Communication (SC), a paradigm poised to optimize data exchange efficiency while maximizing spectrum resource utilization. However, despite notable advancements, current SC systems face challenges related to context comprehension and the integration of prior knowledge [26, 88–96]. GAI emerges as a viable solution, providing the ability to automate various tasks and produce personalized content across diverse communication channels. Moreover, within the architecture encompassing cloud, edge, and mobile computing, Semantic Communication (SC) networks play a crucial role, as highlighted by Prasad and colleagues [97]. The implementation of GAI, both globally and locally within these networks, is instrumental in refining the transmission of semantic content, the synergy of source and channel coding, and the adaptability in generating and encoding information. This approach aims to enhance semantic understanding and optimize resource allocation.

Another study conducted by [98] examined the diverse applications of GPT models in Natural Language Generation (NLG) tasks. Drawing inspiration from OpenAI's groundbreaking GPT-3 innovation, the article provided a comprehensive overview of ChatGPT's current role in NLG applications. It explored potential uses across various sectors including e-commerce, education, entertainment, finance, healthcare, news, and productivity. A crucial aspect of the analysis involved evaluating the efficacy of implementing GPT-enabled NLG technology to customize content for individual users.

In [99], the authors outlined recent advancements in pre-trained language models and their potential to enhance conversational agents. Their analysis focuses on assessing whether these models can effectively address the challenges encountered by dialogue-based expert systems. The article examines various underlying architectures and customization options. Additionally, it explores the existing challenges encountered by Chatbots and similar systems. The emergence of pre-trained models (PTMs) has sparked a revolution in Natural Language Processing (NLP), significantly impacting the field. Qin et al. [100] offered a comprehensive

examination of Pretrained Models (PTMs) in Natural Language Processing (NLP). They provide insights into applying PTM knowledge to downstream tasks and suggest potential research directions for the future. This survey serves as a valuable guide for comprehending, leveraging, and advancing PTMs across various NLP applications.

Authors in [101] delved into the background and technological aspects of GPT, with a specific focus on the GPT model itself. The article explores the utilization of ChatGPT for developing advanced chatbots. Within the study, interview dialogues between users and GPT are showcased across various topics, and subsequently analyzed for semantic structures and patterns. The interview underscores the advantages of ChatGPT, emphasizing its capability to enhance various aspects including search and discovery, reference and information services, cataloging, metadata generation, and content creation. Important ethical issues like bias and privacy are also covered. Another study in [102] discussed the benefits and downsides of using LLMs in the classroom while taking into account the perspectives of both students and teachers.

They place a strong emphasis on the opportunity for creating educational content, raising student engagement, and modifying learning environments. However, the use of LLMs in education has to be backed with effective prompt engineering, to ensure high competency among the students and faculties. Table 3 presents a comparitive study of existing surveys.

1.2 Uniqueness of the Survey

In contrast to previous reviews that predominantly focus on foundational architectures, auxiliary model frameworks, and specific technical details related to GPTs and LLMs, this proposed survey introduces a novel bifurcated methodology. It offers a comprehensive examination of ChatGPT by elucidating its architectural intricacies, delineating its diverse deployment landscapes, and conducting a thorough evaluation of its training methodologies. Thus, the survey provides a holistic perspective on the complex designs and development of ChatGPT-centric models tailored for specific application domains. By juxtaposing ChatGPT with alternative learning frameworks, this study aims to illuminate the distinctive learning mechanisms employed by AI models under various training paradigms and conditions. This addresses a gap in existing literature, which often portrays GPTs solely as conversational agents, by providing empirical support for the utilization of GPT technology in real-time data analysis applications.

The survey provides an in-depth exploration of the pivotal technologies that form the foundation of ChatGPT's functionality. It acts as a bridge between strategic architectural decisions and the tangible aspects of resource distribution and performance enhancement. A noteworthy case study EcoGPT by Chen et al. [103] is examined to highlight the economic considerations in optimizing GPT models, introducing three principal strategies, namely, the LLM approximation, prompt adaptation, and model cascading. These strategies demonstrate significant cost efficiencies, thereby facilitating the deployment of cost-effective models in practical scenarios. Hence, this survey aims to provide both researchers and industry professionals with invaluable insights to develop superior GPT models. These models are not only customized for specific applications but also effectively balance the cost-scalability equation.

1.3 Survey Methodology

In this subsection, we discuss the potential research questions which the authors brainstormed while preparing the survey. Table 4 presents the research questions, and also presents its practical importance (why these RQs needs to be addressed) Based on the proposed RQs, the authors have done a thorough scan of literature works associated with GPT models from academic databases. A list of possible keywords involved in the search include "ChatGPT", "GPT", "LLM + Training", "GPT models", "Security for GPT models", "GPT architecture", "Model training for LLMs", "Prompts in LLMs", "Finetuning LLMs", "Performance of LLM models", "GPT in applications", and others. The search is performed on literary databases like IEEE, Elsevier, Taylor & Francis, Springer, ACM, and Wiley, and also indexing platforms like Google scholar, and preprints like arXiV, bioRxiv, and SSRN were scrutinized.

The initial search yielded a significant pool of approximately 250 academic articles. These articles were subsequently refined using stringent inclusion and exclusion criteria. The first phase involved a thorough review of titles and abstracts, resulting in a reduction to 250 relevant publications. Following this, the introductions of these articles were scrutinized to ensure alignment with our established research topics, leading to a further reduction of 80 publications. The third step involved a meticulous content evaluation tailored to our research objectives, resulting in the identification of 165 papers that met our criteria for academic rigor and relevance. The core of our thorough survey is built on these chosen works.

1.4 Survey Contributions

The research contributions of the article are presented as follows.

- We delve into the fundamentals of Large Language Models (LLMs) to bolster Chat-GPT models, focusing specifically on the self-attention mechanism, Reinforcement

Table 3 Comparative study of existing state-of-the-art approaches

Author [2–14]	Industry Involvement									Technology				Remark
	Technical	Education	Business Service	Manufacturing	Finance	Healthcare	Government	Entertainment	Agriculture	AI	HCI	Big Data	CC	
Mondal et al. [66]	✓	✓	✗	✓	✗	✗	✗	✓	✗	✓	✓		✗	Despite the advance ment in GAI, the technology is still not in its prime due to ethical and practical issues. The use of reinforcement learning still makes incorrect decisions
Cooper et al. [67]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	We are in the early stage of using GAI in education, but ChatGPT is definitely a part of research in education and its implications in the future
Shoja et al. [75]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓		✗	It is good to use GAI for scientific research work due to speculation, but it is important to manage risks associated with professional life that may be exposed to legal issues
Lv et al. [77]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓	✗	Different technologies in the metaverse become the driving force behind future enhancements. Integration of AI with ChatGPT creates new socioeconomic space
Gozalo-Brizuela et al. [78]	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓	✗	Generative AI has created space for creativity and personalization, and this can help to optimize creative and non-creative tasks, but it faces issues with finding the of correct database

Table 3 (continued)

Author 2–14	Industry Involvement									Technology				Remark
	Technical	Education	Business Service	Manufacturing	Finance	Healthcare	Government	Entertainment	Agriculture	AI	HCI	Big Data	CC	
Baidoo-Anu et al. [86]		②		◇	◇	◇		◇	◇	②	◇	◇		Despite several limitations, ChatGPT will revolutionize the education sector. GAI helps teachers to identify any AI-generated assign- ment. But with current NLP AI detectors are not accurate enough
Xia et al. [87]	②	②		◇	◇		◇	◇	◇	②	◇		◇	Explored the potential of AI-generated content service for semantic communication and highquality and efficient content delivery by reducing the transmission traffic
George et al. [98]	◇		②	②	②	◇		◇	◇	②	◇	◇		ChatGPT version 4 explores the potential of business communication in corporate environments to enhance customer service and their experience. And how it alters existing norms in a business setting
Zhang et al. [99]	②	②		◇	◇	◇		②	◇	②	②	◇		Explored the capabilities of GPT-4 in text, images, and video. The GPT models still face many challenges, such as a lack of interpretability, ethical and legal concerns, along with technical challenges that are domain-specific
Qiu et al. [100]	◇	◇	◇	◇	◇	◇	◇	◇	◇	②	◇	◇	◇	Comprehensive on pre-trained models, their architecture and extensions but not provided the industry implementations of these models in general

Table 3 (continued)

Author 2–14	Industry Involvement									Technology				Remark
	Technical	Education	Business Service	Manufacturing	Finance	Healthcare	Government	Entertainment	Agriculture	AI	HCI	Big Data	CC	
Lund et al. [101]		②	Absent	Absent	Absent	Absent	Absent	Absent	Absent	②	Absent	Absent	②	How Chat-GPT helps and impact education sector and discusses the history and model architecture along with its sophisticated working
Kasneci et al. [101]		②	Absent	Absent	Absent	Absent	Absent	Absent	Absent	②	Absent	Absent	Absent	The large language is a key research area in education sector that has created opportunities for researchers. To unleash the full potential it is important to use such models with caution
Proposed	②	②	②	②	②	②	②	②	②	②	②	②	②	–

Industry Application ②: Present   : Partial   : Absent

Table 4 Research questions addressed in the survey and its practical importance

RQ No	Research Question	Practical Significance
RQ1	What learning paradigms and architectural considerations are pivotal in the development and performance of LLMs?	Understanding foundational principles for effective deployment and customization
RQ2	What are the specific architectural and operational details of the ChatGPT model?	Enables informed integration of Chat-GPT into diverse applications
RQ3	What underlying technologies are crucial for supporting the Chat-GPT model, and how do they contribute to its functionality?	Facilitates reliability and scalability in real-world settings
RQ4	What are the open issues and challenges in deploying and maintaining ChatGPT and other LLMs?	Identifies areas for technological improvement and resource allocation
RQ5	What are the security and ethical considerations when utilizing the ChatGPT model in various applications?	Ensures responsible and safe usage in diverse environments
RQ6	What optimization strategies can be employed to design cost-effective GPT-based solutions?	Lowers operational costs while maintaining or improving performance

Learning based Human Feedback (RLHF), and prompt adaptation techniques for GPT models.

- Building upon the foundational concepts of ChatGPT, we elucidate the architectural and functional components of the ChatGPT model. This exploration paves the way for intriguing possibilities in task-specific and training optimizations, fundamentally reshaping the deployment of ChatGPT across various application domains.
- We also examine the underlying technologies that underpin ChatGPT. This explo-ration not only satisfies academic curiosity but also holds practical implications for industries aiming to incorporate Large Language Models (LLMs) into their operations.
- We explore the open issues and challenges associated with deploying Chat-GPT, delving deeper into the security and ethical considerations of GPT models. This discussion lays the groundwork for ensuring the secure utilization of GPT technology.
- Lastly, our case study, EcoGPT, showcases core optimizations aimed at reducing costs and training time for GPT models. Through performance evaluation, we identify practical changes necessary to redesign GPT models for large-scale deployment.

1.5 Layout

The rest of the article is divided into eight sections. Section 2 discusses the preliminaries (nuts and bolts) to design GPT models. Section 3 presents the architectural schematics of the LLMs, and then analyses the ChatGPT model, with discussion on the key AI models. Section Sect. 4 presents the open issues and challenges in GPT model design (in terms of training, resource requirements, and model designs). Section 5 presents the security and ethical considerations to be followed in the GPT design. Section 6 presents the case study, EcoGPT, which renders a cost-effective and fine-tuned LLM design. Section 7 summarizes the key points and

discusses the future directions of research in generative AI and GPT models. Finally, 8 presents the final thoughts and concludes the article with potential future scope of work by authors.

2 The Nuts and Bolts of ChatGPT

With its exceptional performance across a broad spectrum of textual outputs, ranging from scientific publications to creative works, ChatGPT stands at the forefront of AI language models. It offers a reliable approach to content recognition by leveraging an AI classifier trained on an extensive corpus of text generated by both humans and AI. Through its provision of contextually aware and enlightening responses, its advanced Natural Language Processing (NLP) capabilities redefine human-AI interaction. Acting as a versatile tool, ChatGPT fosters innovation and enhances efficiency across various industries. Thus, in this section, we delve into the fundamentals, often referred to as the nuts and bolts, of ChatGPT. This lays the groundwork for readers to gain a deeper appreciation and understanding of the ChatGPT architecture and key components discussed in the subsequent sections.

2.1 Evolution of ChatGPT

On November 30, 2022, OpenAI took a significant stride towards democratizing AI technology by releasing ChatGPT for public testing. In just the first week, user adoption surged, with over 200,000 sign-ups, underscoring the immediate appeal and utility of the model. Following this, on December 8, 2022, ShareGPT was introduced, encouraging community participation through shared conversational experiences. This initiative resulted in a notable 15% increase in user-generated content, according to metrics published by OpenAI. As of February 1, 2023, OpenAI shifted towards

a more business-oriented approach with the introduction of GPT Plus, a subscription-based service. Garnering an initial uptake of 50,000 subscribers, it provided exclusive benefits, positioning it as a premium option for professionals and enterprises. On February 7, 2023, Microsoft broadened the model's scope of application by integrating it into the Bing search engine, resulting in a 20% increase in daily search traffic on Bing. Subsequently, OpenAI facilitated more extensive platform integrations on March 1, 2023. Then, on March 9, 2023, Azure OpenAI Services enabled 300 inter-API platform integrations.

On March 14, 2023, GPT-4 became accessible to Chat-GPT Plus members, enhancing the context accuracy of the GPT-3.5 model by 10%. Just nine days later, on March 23, 2023, plugins were launched, offering a broader range of customization options. Within a month, subscriptions increased by 25% owing to effective user customization, which provided personalized and on-the-fly services. Subsequently, on April 4, 2023, chat-integrated services-focused firms with Y Combinator backing entered the market, stimulating entrepreneurship activity.

On April 20, 2023, Yokosuka, a Japanese city, began utilizing ChatGPT for administrative tasks, showcasing the model's global applicability. The code interpreter plugin, which was launched on July 16, 2023, solidified ChatGPT's status as a versatile tool in the contemporary computing landscape. Figure 1 presents the timeline of ChatGPT's evolution, illustrating the multifaceted progression of ChatGPT.

2.2 Learning mechanisms integrated with ChatGPT

Table 5 indicates about the various learning techniques that can be compared with the ChatGPT learning mechanism. The framework known as Abductive Learning (ABL) [104], which aims to link ML and logical reasoning. ABL uses a knowledge base that makes use of logical languages like First-Order Logic (FOL), in contrast to traditional supervised learning techniques. Conclusions derived from knowledge-based reasoning offer alternatives to conventional supervised labels. Abductive learning integrates a learning model, such as a neural network, capable of learning from both labeled and unlabeled datasets. This approach embodies a standard process guided by data induction. In contrast, reasoning occurs when predictions from the learning model are applied, benefiting from knowledge-based data. The effectiveness of the learning model is affirmed when there is alignment between the outcomes of the reasoning and the information within the knowledge base. The process entails iteratively executing two steps: induction and deduction, which are akin to learning and reasoning, respectively. These steps persist until logical reasoning, based on the predictions made by the learner, coincides with the knowledge base. However, if achieving consistency becomes unattainable, the

learning process is halted and terminated. Another training method known as "parallel learning" involves running several models or learning tasks at once [105]. Parallel learning enables the simultaneous training of many models or learning tasks as opposed to a single model sequentially, which may result in a faster and more effective learning process. This strategy proves particularly beneficial in scenarios with substantial volumes of data. In such cases, learning tasks can be subdivided into smaller components and executed concurrently. To optimize learning effectiveness and leverage computational resources efficiently, parallel learning can be achieved through distributed computing resources or parallel processing algorithms. The parallel learning framework typically consists of three modules: a description module that reconstructs the real world into artificial worlds; a prediction module that runs computational experiments with the learner in the artificial worlds; and a prescription module that applies the prediction (action) to the real world in order to manage and control it.

There is another learning technique which is called as Federated Learning, McMa-han et al. [115] first put up the idea of Federated Learning (FL) in 2016. By creating a decentralized federation, where numerous clients only communicate with a central server, FL aims to preserve data confidentiality and privacy. Each client has a unique dataset that is kept private from other clients. The selected group of clients receives a global model from the server to use as a starting point for training on their individual local datasets. The server then compiles the locally trained models from each of the chosen clients and updates the global model using an average operation. Throughout the FL process, these two steps—model distribution and aggregation—are repeated iteratively. In order to improve machine translation another learning technique was identified and is called dual learning, and was first implemented in 2016 [107]. The primary objective of dual learning is to address the challenge of acquiring extensive labeled data. Unlike active learning, which depends on manual labeling, dual learning leverages the structural resemblance between two tasks, maximizes the utilization of unlabeled data, and enables bidirectional enhancements in learning performance. In the context of dual learning, these two machine learning tasks exhibit structural duality, where one task involves mapping from space X to space Y, and the second task involves mapping from space Y to space X. This inherent duality can be advantageous for various applications, such as Neural Machine Translation (NMT) between two languages, Automatic Speech Recognition (ASR) versus Text-To-Speech (TTS), and picture captioning versus image generation.

Bengio first suggested the idea of curriculum learning in 2009 [107]. The curriculum learning can be explained using an example, where the educational material is constantly arranged in a sequential fashion from elementary school

Table 5 Comparison of the ChatGPT with other learning techniques

Learning Mechanism	Year	Learning Motivation	Issues	Necessity	Importance of this mechanism
ChatGPT 2022	2022	To continuously improve the abilities to assist and communicate with humans in a more natural and efficient way	Generating biased or inaccurate responses due to the limitations of the training data and algorithms	Enhance human communication and productivity through advanced natural language processing and ML techniques	Advancing research in NLP and AI
Abductive Learning [104]	2018	Restriction of ML based on the available dataset	incorporating domain-specific knowledge into the implementation process	Humans have a greater capacity to learn through the application of reasoning	The ability to understand and adapt to different contexts and situations
Parallel Learning [105]	2017	Shortage of labeled data leads to safety concerns	Quite challenging with its execution	The virtual world can aid in resolving actual issues	Enhanced precision, expandability, and resilience
Federated Learning [106]	2016	Safeguarding of data confidentiality and personal privacy	Challenges with executing effective communication	Working together in a group with individual independence	Confidentiality, expandability, and efficiency
Dual Learning [107]	2016	Insufficient amount of labeling for a specific task	Constraints on the usage of an application	The coexistence and interdependence of two tasks	Efficient in bidirectional learning with limited resources
Curriculum Learning [108]	2009	The learner's performance is inferior when dealing with complicated tasks	The extent to which the curriculum design's complexity affects the outcome is challenging	User learn gradually, starting with simple concepts and moving towards more challenging ones	fast convergence, improved accuracy, and generalization
Multitask Learning [109]	1997	The learner's performance is inferior when dealing with multiple tasks	Complication, task conflict, and incongruent distribution	Humans learn more quickly on diverse yet interrelated tasks	Increased effectiveness, wider applicability, and decreased overfitting
Transfer Learning [110]	1995	Inadequate labeling for certain tasks	Detrimental transfer, in-congruent distribution, and model prejudice	human uses learned, knowledge to solve, and new tasks	Decreased data requirement, shorter training duration, and enhanced performance
Meta Learning [111]	1995	The learner's performance is inferior when dealing with diverse tasks	Complication, restricted applicability, and in-congruent distribution	Learner possess a meta-skill of abstraction to tackle various tasks	Flexibility, reduced data requirement, and comprehensibility Lifelong
Lifelong Learning [112]	1995	The learner tends to forget prior knowledge as the number of tasks increases	Complication, data retention, and unexpected alterations	Humans can acquire knowledge throughout their lifespan	Shorter training duration, expandability, and flexibility
Ensemble Learning [113]	1990	Challenging to develop a competent learner	Complication, overoptimization, and comprehensibility	Several sub-optimal learners achieve superior performance	Enhanced precision, resilience, and broader applicability
Contrastive Learning [114]	1990	Challenge of preprocessing unlabeled data	Data quality is computationally expensive	Concentrate solely on the difference between samples	Unsupervised impression recital

to university, beginning with simpler topics and moving to more difficult ones. Students who have a firm grasp of the fundamentals typically learn more complex material more quickly and effectively. This finding served as the inspiration for Curriculum Learning, which allows learners to start with simpler samples and work their way up to more challenging ones. The learning process in Curriculum Learning is broken down into T distinct stages, designated as $\langle Q1,..., Qt,..., QT\rangle$ with each step specifying the unique learning criteria for that phase. Here, the Qt is the learning criteria for the learning stage. The criterion in curriculum Learning may include a number of ML-related con-cepts, like as samples, tasks, model capacity, loss, and others. Most curriculum learning frameworks have two components called Difficulty Measurer and Training Scheduler to ensure that each stage, Qt, proceeds from easy to tough. These elements help determine the degree of difficulty of the learning content so that the training process can be scheduled appropriately.

Each sample in the dataset must be given a priority rating by the difficulty measurer in order to determine the order in which they should be learned. The Training Scheduler, meanwhile, decides when and how many difficult examples will be delivered into the learning model. The Difficulty Measurer's main responsibility is to rate the difficulty of components, such as samples. A number of ways can be used to carry out this evaluation, including manual configuration prior to learning, assessment by an external teacher model, dynamic evaluation depending on the model's output, or even reinforcement learning methods. A useful strategy that can improve the performance, robustness, and convergence speed of ML models is curriculum learning. The caliber of the curriculum's design and the dataset being used, however, determine how effectively students learn from a curriculum. To improve the learner's capacity for generalization, Caruana [109] created Multi-Task Learning (MTL) in 1997. This is accomplished by utilising domain-specific data from related task training data. MTL can be viewed as an inductive transfer mechanism in which the training signal from a second task acts as an inductive bias to speed up learning. The NeurIPS workshop "Learning to Learn: Knowledge Consolidation and Transfer in Inductive Systems" in 1995 helped popularise the Pratt [110] notion of transfer learning. The observation that people may use the knowledge they've learned to solve new, related problems more quickly and effectively is what drives transfer learning. Transfer learning seeks to solve issues with less labeled data by loosening the assumption of independent and identically distributed (i.i.d.) training and test data. Its goal is to transfer the information learned from a source domain—such as task Ts on dataset Ds —to a target domain (task Tt and dataset Dt), where there is a lack of labeled data. The final objective is to increase the prediction function's ft learning performance in the intended domain. Transfer learning can be seen as an extension of

Meta learning [111], which is a subset of the more general idea of Learning to Learn. It involves using many source tasks (T1, T2,..., Tn) to help solve a new task, Tt, more quickly. Di, the dataset for each task, has two subsets: D^{tr} for training and D^{ts} for testing. Depending on how the meta-knowledge is learned and applied to new activities, there are three different types of meta-learning. Model-based, optimization-based, and metric-based approaches are some examples of these categories.

Thrun and Mitchell [112] first proposed the idea of lifelong learning in 1995. It is based on the notion that the experience and knowledge we acquire via ongoing learning throughout our lifetimes can improve our capacity to meet new difficulties in the future. It is because of this realization that lifelong learning is being developed. Lifelong learning entails a continuous and incremental learning process where the learner has completed a number of tasks (T1, T2,..., TN) at a specific period. These tasks may fall under a variety of domains and are linked to the appropriate datasets (D1, D2,..., DN). A knowledge base is used to store and make use of the knowledge obtained from these tasks. The goal of lifelong learning is to use the knowledge that has been gained within the KB to help complete a new task (TN + 1), then to update the KB with the newly acquired knowledge. In conclusion, lifelong learning demonstrates a number of critical features, including the following: (1) it is a continual process of learning; (2) it involves the collection and updating of knowledge; and (3) it makes use of the collected knowledge to support new learning activities. Deep neural network learning's catastrophic forgetting problem has been addressed via lifelong learning, and learning in open environments has also been studied. Lifelong learning has found use in the field of NLP, helping to improve topic modelling, information extraction, and chatbots' ability to engage in continuous conversation.

A ML technique known as ensemble learning combines the results of various distinct models to produce a single, more reliable and accurate prediction. Ensemble learning [113] makes use of the variety and combined knowledge of numerous models, as opposed to relying on just one, to enhance overall performance. Ensemble learning involves training multiple base models on the same dataset using various techniques or data sampling alterations. The ensemble method merges the predictions from each individual model in a meaningful manner to generate the final forecast. This combination can be achieved through techniques such as voting, averaging, or weighted averaging. Ensemble learning offers several advantages, including improved generalization and reduced overfitting by averaging the predictions from multiple models to mitigate the impact of individual model biases or errors. Additionally, it might be able to capture many facets of the data and offer a more thorough insight of the underlying trends.

Ensemble learning encompasses several effective techniques, including bagging, boosting, and stacking. Bagging involves combining the outputs from multiple models, with each model trained on different bootstrap samples drawn from the training dataset. In contrast, boosting adopts a sequential approach to model training, where each subsequent model is tailored to correct the errors of its predecessor. Stacking, on the other hand, utilizes a meta-model that optimally weighs the inputs from various base models to generate a consolidated prediction. With the advent and subsequent success of large-scale pretrained models in the realms of Natural Language Processing (NLP) and Computer Vision (CV), the concept of contrastive learning, which has its roots in the work by Hinton in the early 1990s [114], has witnessed a resurgence in popularity, showcasing its relevance and applicability in the modern era of deep learning advancements.

Contrastive Learning has emerged as a crucial tool in unsupervised scenarios, functioning as a form of self-supervised learning. Its core tenet posits that similar examples should exhibit comparable representations, with the learner focusing on discerning distinctive features rather than mastering every aspect of the sample. The objective of contrastive learning is to separate the representations of dissimilar examples while bringing together those of similar examples. This is achieved through the design of a contrastive loss function and appropriate selection of metrics. Contrastive learning has been effectively applied in various tasks, including picture-to-image translation, image captioning, autonomous driving, and visual positioning.

2.3 Transformers: The ChatGPT base

The base architecture of ChatGPT model is the transformer model [8], which is an encoder-decoder structure. The designed models on transformers are trained on billion of parameters, which are systematically arranged into layers and head units. The encoder-decoder fundamentally have several layers, where each layers consists of multiple head units to focus on different parts of the input text. Figure 4 presents the transformer architecture.

2.4 A Round of Encoding Operation

Mathematically, an encoder works on an input sequence $X = \{\times 1, \times 2, ..., xn\}$, where n is the sequence length. Every x_i is a d-dimensional input embedding. The input embedding is shown mathematically as follows.

$$E = X \cdot We \qquad (1)$$

where We is the embedding-weight matrix. Next, we have positional encoding P added to the input embeddings E, which is represented as E′, and thus E = E + P. Next, stage

is the multi-head self-attention mechanism, where E′ is linearly projected into a query Q, key K, and a value V. The tuple (Q, K, V) are matrices, and are represented as follows.

$$Q = E\prime \cdot W_q, \ K = E\prime \cdot W_k, \ V = E\prime \cdot W_v \qquad (2)$$

The values are split into h different heads and scaled dot product attention H_i on each head is computed as follows

$$H_i = \text{Attention}(Q_i, K_i, V_i) \qquad (3)$$

These attention heads are concatenated and one more time are linearly transformed, which results in the multi-head attention output, presented as follows.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, H_2, ..., H_h) \cdot W^O \qquad (4)$$

Once the multi-head self-attention layer is complete, the next step in the encoding process if the Feed-Forward Neural Network (FFN), which is applied position wise in independent fashion to the output. The FFN is basically two linear layers with a Rectified Linear Unit (ReLU) activation in between, presented as follows.

$$\text{FFN}(x) = \text{ReLU}(x \cdot W_1 + b_1) \cdot W_2 + b_2 \qquad (5)$$

Post the FFN, layer normalization is done, and the processed sequence is passed to the next encoder layer, which applies the same steps of operation. The following encoding process is successively repeated for all encoding layers.

2.4.1 A Round of Decoding Operation

The Decoder also consists of layers of multi-head attention and feed-forward neural networks but has an additional layer of multi-head attention connecting it to the output of the Encoder. First, we have the Masked Multi-Head Self-Attention, which is similar to the Multi-Head Self-Attention but with a mask to prevent future positions from being attended to. After this layer, we have the encoder-decoder attention layer, which takes the output Z of the last Encoder layer and the output Y of the Masked Multi-Head Self-Attention layer of the Decoder. The details are as follows.

$$\text{Multi Head}(Q, K, V) = \text{Concat}(H_1, H_2, ..., H_h) \cdot W^O \qquad (6)$$

where $Q = Y \cdot Wq$, $K = Z \cdot Wk$, $V = Z \cdot Wv$.

At the final layer, the decoder output goes through a linear layer followed by Softmax layer for prediction. It is shown as follows.

$$\text{Output} = \text{Softmax}(Y \cdot W) \qquad (7)$$

The data used for training these models usually consist of large text corpora that contains websites, books, and articles. These are tokenized into subwords or words, followed by embedding into continuous vectors. The model is

Fig. 4 Transformer Architecture
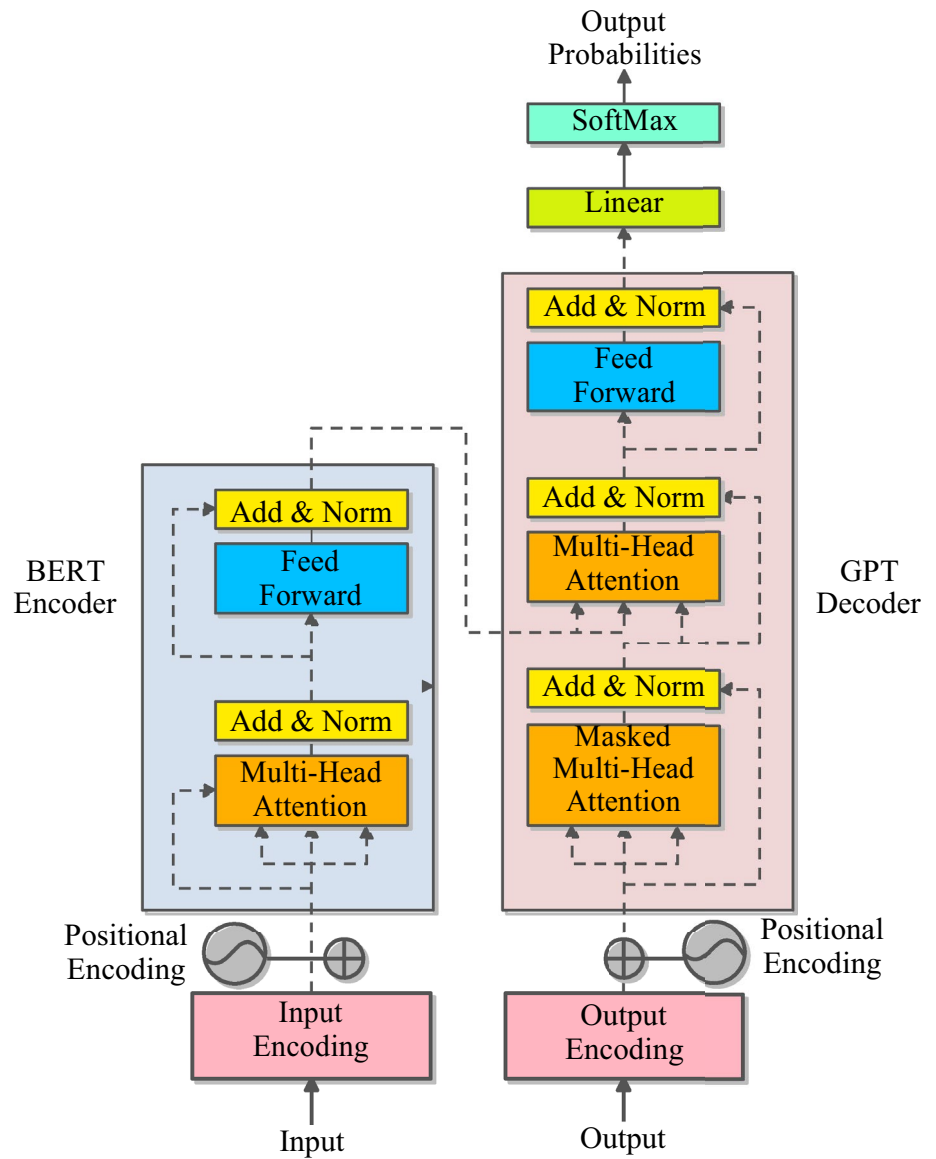
trained using optimization algorithms like Adam with custom learning rates, often employing techniques like learning rate warm-up and layer normalization.

2.4.2 Tokenization and Embedding

Tokenization segments a sequence of text into discrete units known as "tokens." These tokens represent words, subwords, or single characters, depending on the specific tokenization approach chosen. This can be defined as:

Input: A textual sequence, denoted as $T = \{w1, w2,..., wn\}$, where w1, w2,..., wn are the generated words by the token.

Output: A list of tokens, represented as tokens $= \{t1, t2,..., tm\}$, with 'm' indicating the number of tokens generated

through the tokenization process. The token creation is shown as follows.

$$Ttoken \leftarrow (t1, t2, ..., tm) \tag{8}$$

Initially, T consisted of words w1, w2,..., wn, and the tokenization process converts each word into the number of tokens. The tokenization splits the text based on the white space or punctuation and employs a byte-pair encoding technique to convert the words into tokes. Word embedding represents words as continuous vectors in a multi-dimensional space while maintaining the semantic relationships between words.

$$Wembedding \leftarrow E(T_{token}) \tag{9}$$

where $E(T_{token})$ is a high dimensional vector representing semantic information about the tokens, this can be done using cosine similarities and an unsupervised learning algorithm. Two integrated techniques, the self-attention mechanism, and the reinforcement learning strategies to the encoding–decoding process is discussed in next section. Also, ChatGPT allows a rich set of Application Programming Interfaces (APIs) to allow interoperability with different applications, which makes it a gamechanger to connect with open-source LLM designs be developers. The details are discussed in the next subsection.

2.5 APIs and Interoperability

API is a set of rules that allows one software to interact with another application. API makes it easier to utilize the functionality of any application without having implementation knowledge by interacting with servers and back-end systems. The ChatGPT API is powered by the "gpt-3.5-turbo" model, representing OpenAI's most advanced language model available for public use, devoid of any subscription requirements. This model features an expansive context window of 4096 tokens, marking an enhancement of 96 tokens compared to its predecessor, the Davinci-003 model. Furthermore, there exists a "snapshot" iteration named "gpt-3.5-turbo-0301," notable for its exemption from automatic updates. This version grants developers the autonomy to integrate OpenAI's latest model enhancements into their systems at their discretion. OpenAI is committed to the continuous refinement of the ChatGPT API, with scheduled updates and improvements slated to commence from April 2023, ensuring the API remains at the forefront of technological advancements.

The OpenAI GPT-3.5-Turbo API offers developers a multitude of advantages for enriching their applications with natural language processing (NLP) capabilities. One notable advantage is its capacity to generate text responses that closely resemble human language, thereby elevating the overall user experience of these applications. This capability unlocks avenues for creating diverse content, including blog articles, social media posts, and more. Moreover, the OpenAI GPT-3.5-Turbo API is distinguished by its continuous learning and improvement. As the model interacts with more data, its accuracy and efficacy consistently improve. This iterative learning process ensures that developers can depend on the API to provide responses that are not only up-to-date but also highly precise when addressing user inquiries. This dynamic aspect of the API presents a promising trajectory for the advancement of language-centric applications.

2.6 Key enablers for ChatGPT

With cutting-edge developments like Big Data, AI, Cloud Computing, enhanced networking like 5G and beyond, and Human–Computer Interaction, GPT is the culmination of numerous technical advancements [116]. Figure 5 provides an overview of the technologies that enable GPT working with rich interfacing in diverse applications. GPT provides the representation of the essential elements of the GPT models.

- Bigdata: Big data is the term used to describe the enormous amounts of structured and unstructured data that are produced by organizations, people, and machines. The emergence of new technologies, like the Internet of Things (IoT) [117], has sparked an explosion in the production of data from sources like social media, sensors, and transaction-based systems. The advent of big data has revolutionized how businesses approach data analysis and decision-making. The training it has facilitated for the utilization of advanced models like GPT in the field of natural language processing (NLP) has yielded invaluable insights. GPT models, in particular, leverage deep learning (DL) techniques and big data for natural language creation. They harness vast amounts of data, ranging from millions to even trillions of samples sourced from diverse repositories such as books, papers, websites, and social media platforms.

Fig. 5 Supporting technologies for the ChatGPT
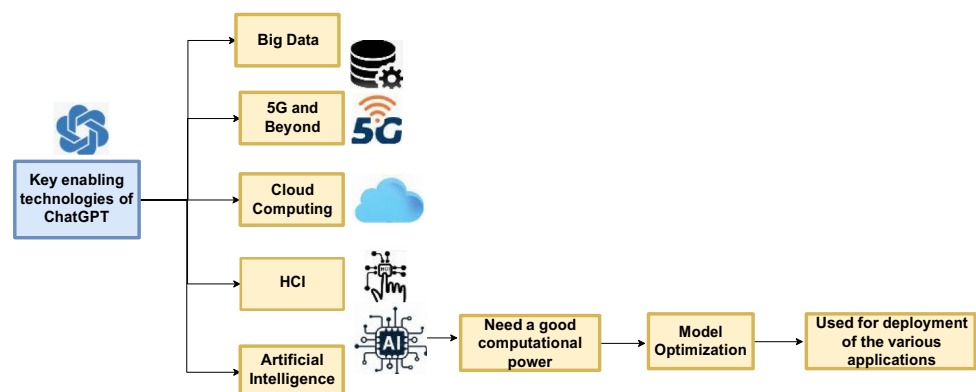
- An exemplar of this advancement is the sophisticated and multimodal GPT-4. With the aid of extensive and diverse training data, GPT models, including GPT-4, exhibit heightened accuracy and proficiency in NLP tasks such as questionanswering, text summarization, and language translation. Furthermore, their adaptable and versatile nature allows for customization to specific tasks and domains, and they can even be extended to produce photos and videos.

Big data offers numerous advantages for GPT, such as the capability to train with vast amounts of data. However, it also presents challenges concerning data accuracy, privacy concerns, and ethical data usage. Nevertheless, with the ever-growing accessibility of data, GPT models are poised to advance significantly, heralding a transformative future for NLP. To harness the full potential of big data and GPT models, organizations must prioritize addressing ethical concerns and ensuring data accuracy as technology progresses.

- Human–Computer Interaction: In the realm of Human–Computer Interaction (HCI), ChatGPT benefits from several key enabling technologies that have significantly augmented its capabilities and improved user experience. Among these, Natural Language Processing (NLP) techniques stand out as fundamental pillars. Transformer-based models, in particular, have played a pivotal role in ChatGPT's success within HCI. These models leverage attention mechanisms to grasp contextual nuances, resulting in more coherent and contextually relevant responses that mimic human conversation. Moreover, advancements in pre-training techniques have been instrumental. Techniques such as unsupervised learning and large-scale language modeling have empowered ChatGPT to generalize effectively across diverse domains. This versatility makes ChatGPT a valuable tool for various HCI applications, enhancing its adaptability and utility in real-world scenarios.

Another critical technology for ChatGPT (GPT-4 and above) is the integration of multimodal capabilities. Incorporating multiple modalities such as images, sounds, and videos alongside text can greatly enrich user interactions. ChatGPT leverages techniques from computer vision and audio processing to comprehend and generate content across different modalities, resulting in more dynamic and immersive user experiences. This multimodal approach is particularly relevant in Human–Computer Interaction (HCI), especially in the context of virtual assistants. These assistants cater to user interactions through a combination of voice commands, text inputs, and visual elements, making the ability to handle multiple modalities invaluable for enhancing user engagement and satisfaction. Overall, these pivotal technologies, grounded in advancements in Natural Language Processing (NLP) and multimodal integration, have positioned ChatGPT as a robust and adaptable tool for improving user experiences across diverse Human–Computer Interaction (HCI) scenarios.

- Cloud Computing: In the context of Cloud Computing, key enabling technologies for ChatGPT play a critical role in harnessing the potential of AI and NLP on a scalable and efficient cloud-based infrastructure. One essential component is cloudbased infrastructure [118], which provides the computing resources required for training and deploying large-scale language models like ChatGPT, such as Graphics Processing Units (GPUs), and Tensor Processing Units (TPUs). Cloud platforms offer elasticity, enabling enterprises to allocate resources dynamically as required, facilitating the execution of compute-intensive tasks such as training and deploying advanced AI models like ChatGPT. Additionally, containerization and container orchestration technologies like Docker and Kubernetes are essential for deploying ChatGPT on the cloud [119].

Containers serve to encapsulate the program and its dependencies, ensuring consistent and repeatable deployments across multiple cloud instances. Kubernetes, functioning as an orchestration platform, automates container deployment, scaling, and management, facilitating the seamless and efficient operation of ChatGPT at scale in cloud environments. When coupled with robust cloud-based AI services, these technologies establish a foundation for integrating ChatGPT into cloud computing ecosystems. This integration enables a diverse array of applications, spanning from chatbots to natural language interfaces for cloud-based services and resources [120]. Cloud providers utilize recommender systems to optimize resource provisioning and service allocation for models like ChatGPT. These systems analyze usage patterns, workload characteristics, and performance metrics to suggest appropriate resource allocations, instance types, and configurations. This proactive approach enhances system reliability, scalability, and cost efficiency, ensuring that ChatGPT receives the necessary resources and services to operate effectively within the cloud environment [121].

- 5G and beyond communication networks: The emergence of 5G and 6G technologies presents significant potential for augmenting the functionalities of expansive language models like GPT-3.5 and GPT-4 [122]. The ultra-fast data transfer capabilities of 5G technology provide uninterrupted access to cloud-based models [123], resulting in expedited retrieval of information and enhanced real-time language processing on user devices [124]. The

reduced latency observed in these networks also facilitates the lightning response of the model, which enhances the user experience in applications. The integration of 6G networks provides the opportunity for localized language model processing on user devices, hence ensuring privacy and reducing on cloud infrastructure [125, 126]. The aforementioned technical improvements have the potential to expand worldwide accessibility to LLM and enhance the availability of advanced language processing skills. The improved data transfer rates and reduced latency associated with 6G technology contribute to enhanced efficiency in model training, ultimately refining the output response to queries. Moreover, the integration of advanced AI and ML algorithms in 6G networks enables models to access more sophisticated algorithms, leading to more refined and contextually appropriate responses. This transformative capability has the potential to revolutionize interactions with AI across various domains. Additionally, to bolster data privacy, federated learning techniques are integrated with 5G networks. [127].

- Artificial Intelligence: AI resides at the heart of ChatGPT LLMs, and it plays an essential role in these models and provides services in all area, including healthcare [128] [129], military surveillance [130], and many others. Understanding and producing human-like language necessitates the use of advanced ML algorithms and neural network topologies. ChatGPT can evaluate massive volumes of text data, learn linguistic patterns, and provide contextually relevant responses thanks to AI, particularly DL. Additionally, AI fuels ChatGPT's ongoing evolution through finetuning and transfer learning, enabling its versatility across diverse domains and applications. Nevertheless, fully unleashing AI's potential in Large Language Models (LLMs) like ChatGPT demands substantial processing capacity. The sheer scale and intricacy of these models necessitate robust processing capabilities. LLMs boasting billions of parameters rely on access to high-performance GPUs or TPUs for effective training and fine-tuning. To mitigate processing demands during inference and deployment, model optimization techniques like quantization and pruning come into play. These AI-driven optimizations enhance ChatGPT's resource efficiency while preserving its linguistic proficiency. Moreover, employing ChatGPT across various applications such as virtual assistants, content generation, and language translation underscores the need for scalable cloud-based infrastructure [131]. AI-driven solutions such as auto-scaling and containerization guarantee seamless deployment of ChatGPT across multiple applications, efficiently utilizing computational resources as required. In summary, AI empowers ChatGPT LLMs to excel in natural language understanding and generation. However, effective utilization mandates access to substantial computational resources, meticulous model optimization, and flexible deployment strategies tailored to diverse applications.

3 Demystifying ChatGPT: Insights into OpenAI LLMs

The advent of Large Language Models (LLMs) has ushered in a new era of Natural Language Processing (NLP) and comprehension, with ChatGPT emerging as a prominent exemplar. This study aims to delve into the intricacies of OpenAI's LLMs, shedding light on their underlying mechanisms, capabilities, and implications. ChatGPT's development marks a significant milestone in AI research, solidifying Ope-nAI's commitment to advancing AI in a responsible and ethical manner. Through an exhaustive examination of its architectural design, training methodologies, and ethical considerations, this research endeavors to provide a comprehensive resource for scholars, practitioners, and policymakers alike. One of the primary distinguishing features of ChatGPT is its challenging scale and sophistication. With billions of parameters, it demonstrates remarkable prowess in generating coherent and contextually relevant natural language responses. This study will dissect the underlying architecture responsible for such capabilities, with a particular focus on deep neural networks and Transformer-based models. Furthermore, ethical aspects surrounding ChatGPT's utilization will be scrutinized, encompassing bias mitigation strategies, content generation guidelines, and the ethical deployment of LLMs in real-world scenarios. This section aims to contribute to the ongoing discourse regarding the responsible development and deployment of advanced language models by elucidating the inner mechanisms of ChatGPT and tackling the inherent challenges.

3.1 Unveiling the Multifaceted Roles of ChatGPT: From Assistant to Innovator

The comprehensive functions of GPT are depicted in Table 6, which intricately delineates its various roles, its corresponding descriptions, and the pertinent user demographics it serves.

ChatGPT plays a versatile role, serving a diverse user base that includes students, learners, content creators/developers, and general information seekers. It excels in providing clear explanations for complex concepts, facilitating learning and content creation. Researchers leveraging ChatGPT's insights should verify its content against established literature and evidence to ensure accuracy and impartiality in their research papers. Additionally, ChatGPT offers robust support across customer service, education,

Table 6 The prominent functions of ChatGPT in the present situation

Roles of Chat-GPT	Description	Users
Diverse range of linguistic inputs [132]	ChatGPT delivers clear explanations of complex concepts in a variety of speaking tenors while also comprehending a wide range of languages. It can produce text like people and analyze code because it was taught on a large collection of internet content	students, learners, content creators/developers, and General users seeking information
Acquiring the most recent information [133]	ChatGPT adapts quickly to new information, excelling in various fields like customer service, education, and entertainment. It revolutionizes occupations by generating unique content and making certain roles more accessible through assistant-like tools	General users seeking up-to-date information, businesses and organizations utilizing ChatGPT for customer service and engagement, professionals in languagerelated fields, and individuals seeking accessibility and assis tance in various domains
Acquiringknowledge and enhancing skills [134]	ChatGPT's capability to learn from user interactions, enhance its reactions, adapt to various use cases, and comprehend a wide range of subjects gives it a competitive edge, making it a powerful tool for conversational AI development while providing contextually suitable replies	General users seeking improved conversational experiences, AI researchers and developers working on conversational AI systems, OpenAI and its competitors in the AI market, and content creators/data providers who contribute to the training data used by ChatGPT
Beneficial for code creation and recommendations [135]	ChatGPT is versatile, assisting in diverse tasks like code creation, meal recommendations, improving quality of life, unique assignment completion, reacting to cues, providing guidance, acting as a humanoid assistant, and revolutionizing software development	software developers, programmers, general users seeking assistance and guidance, researchers, scholars, and technology sector professionals
Addressing queries [136]	ChatGPTintegratesintobusinesses, swiftly addressing client inquiries, making recommendations, predicting outcomes, aiding code correction, fostering innovation in personalized bots, and supporting contentmoderationonsocialmedia platforms	businesses,customersupport teams,consumers,software developers, and moderators/administratorsofdecentralized social media sites
Utilizations in the businesscontext [137]	ChatGPT's business applications include enhancing marketing campaigns, engaging with target audiences, and achieving marketing objectives. Caution is important to protect private information, and policymakers should consider the risks associated with AI systems	businesses,marketers,policymakers, regulators, employees, users within organizations, and the general public who may experience the impact of ChatGPT and generative AI technologies in business applications
Valuable for professionals in the field of digital marketing [138]	ChatGPT assists digital marketers in enhancing campaigns and engaging with target consumers by creating content for social media updates, blogs, and more. It recommends headlines, opening words, and paragraphs based on keywords	digitalmarketers,marketing teams and professionals, customer service representatives, andmarketresearchers/data analysts
Interpret and conveyideas/notions [139]	ChatGPT creates code, translates concepts from English into programming language, and verifies programmers' language for flaws. ChatGPT resembles human writing, performs NLP tasks, and is trained on a large internet text dataset	Programmers, developers, students, learners, and individuals with an interest in programming and NLP
Improved comprehensibility/understanding [140]	ChatGPT generates code, translates concepts, and verifies programming language. Its popularity lies in its approachability and ability to create human-like writing. Trained on a large internet text dataset, ChatGPT performs NLP tasks	chatbot users seeking improved interpretability,researchers and developers exploring interpretability in AI systems, and marketresearchers/dataanalystsutilizingChatGPTto enhancesurveydesignand extract valuable insights from consumer data

Table 6 (continued)

Roles of Chat-GPT	Description	Users
Superior interpretability [141]	ChatGPT improves interpretation skills through user feedback, enhancing response quality and user experience. It can be used to design intelligent surveys, obtain accurate data, and identify patterns in consumer views and behavior by analyzing customer feedback and unstructured data	general users seeking realistic conversations, researchers and developers in the AI field, technology enthusiasts, and content creators exploring new possibilities in dialogue generation
Offering a dynamic and captivating approach [142]	ChatGPT enhances learners writing by making it dynamic and interesting. It offers customized assignments, real-time prompts, and guidance, generating personalized prompts based on students' interests and abilities	Students, educators, and educational institutions seeking to enhance writing skills and create a more dynamic and engaging writing experience with the assistance of AI technologies
Education [143]	ChatGPT's impressive language understanding benefits education, transforming student connections, aiding teaching, serving as a user-centric search engine, and contributing to writing skill development	students, instructors, and individuals who prefer a conversational search experience, all benefiting from ChatGPT's capabilities in education and its potential to transform how information is accessed and shared
Enhancing and strengthening writing skills [144]	ChatGPT strengthens writing abilities by shifting focus to development, offering precise AI evaluation and feedback, while aiding digital marketers with practical, attractive ad copy generation for increased efficiency	Digital marketers (Professionals in the field of marketing), Writers and content creators, Students, Educators and instructors
Generate innovative concepts or produce novel creations [145]	Generative AI systems like ChatGPT use prior knowledge to create new things through ML learning. They have the potential to reshape employment and provide valuable insights for market research and relationship development	Knowledge workers, researchers and developers, content creators, market researchers, and general users who needs intelligent responses
Encompass a broad spectrum of subjects [146]	OpenAI's ChatGPT covers various topics and writing styles, generating AI responses based on text prediction and extensive training on online human writing	Creative enthusiasts (like artists, poets, and individuals interested in creative writing), writers and content creators, researchers and academics, and students
Perform regular tasks at the office level [147]	ChatGPT automates routine office tasks, optimizes workflow, and enhances cybersecurity with virtual assistants and generative AI technology	Office personnel, employees who require assistance with their day-to-day responsibilities

and entertainment domains, showcasing rapid adaptability to emerging information.

Equipped with assistant-like capabilities, ChatGPT stands poised to revolutionize industries by generating authentic content and making certain professions more accessible. This advancement serves a broad spectrum of users, including individuals seeking up-to-date information, businesses leveraging ChatGPT for enhanced customer interaction, linguists, and those seeking increased accessibility and guidance across various domains and platforms. The content it provides upholds scholarly integrity, ensuring the absence of plagiarism or AI-induced bias, which resonates well with researchers. ChatGPT excels at enhancing knowledge and capabilities. It holds a competitive edge in conversational AI development by its capacity to learn from interactions, adapt, and comprehend diverse topics, delivering contextually appropriate responses. This versatility offers advantages to various stakeholders, including individuals desiring enhanced conversational experiences, conversational AI researchers and developers, industry leaders like OpenAI and its rivals, as well as content creators and data providers enriching ChatGPT's training data. Such capabilities cater to a wide array of needs and interests within the AI landscape, fostering innovation and progress.

In addition to these, ChatGPT boasts advantages across a myriad of other tasks. Its versatility enables it to undertake a diverse range of roles, including code generation, meal suggestions, enhancing overall quality of life, completing creative tasks, handling user inquiries, offering guidance, serving as a virtual assistant, and transforming software development. Such versatility proves beneficial for software developers, programmers, everyday users seeking assistance and advice, researchers, academics, and industry professionals within the technology sector. [148].

To effectively address customer inquiries, offer recommendations, predict outcomes, assist in code enhancement, foster innovation through personalized bot creation, and facilitate content moderation on social media platforms, ChatGPT assumes a pivotal role in query resolution, seamlessly integrating into business workflows. This feature serves a wide array of stakeholders, including companies, customer service teams, end users, software developers, social media moderators, and administrators, benefiting from its capabilities. In corporate settings, ChatGPT finds practical application in supporting marketing endeavors, fostering audience engagement, and achieving overarching marketing objectives. However, safeguarding sensitive data against potential breaches is paramount, necessitating careful consideration of associated risks by policymakers. Enterprises, marketers, regulatory bodies, employees, organizational users, and the general public are among the various groups poised to reap the benefits of such applications. These entities frequently encounter ChatGPT and generative

AI technologies within the realm of business operations. For professionals seeking to enhance their proficiency in digital marketing, ChatGPT proves invaluable, assisting in the creation of content suitable for social media updates, blogs, and other online publications, thereby augmenting campaigns and promoting engagement with target audiences. Moreover, it provides recommendations for tailoring headlines, introductions, and paragraphs to specific keywords. Customer service representatives, seasoned industry specialists, digital marketers, comprehensive marketing teams, diligent market researchers, and data analysts are just a few examples of individuals poised to leverage this technology.

ChatGPT undertakes a multitude of tasks aimed at interpreting and effectively conveying thoughts and concepts. It excels in generating code, facilitating the translation of ideas from natural language to programming languages, and assisting programmers in debugging their code. With capabilities spanning a wide spectrum of natural language processing (NLP) activities, ChatGPT produces outputs comparable to those written by humans, owing to its extensive training on a diverse dataset sourced from the internet. Programmers, developers, students, learners, and anyone interested in programming and NLP stand to benefit significantly from these functionalities. As a versatile tool, ChatGPT offers functionalities such as programming language validation, concept translation, and code generation. ChatGPT's appeal rests on its user-friendliness and its ability to generate text that closely resembles human writing. Leveraging a vast dataset of internet content, ChatGPT adeptly handles a myriad of natural language processing (NLP) tasks. Its utility extends to various user groups, including enthusiasts seeking improved interpretability in chatbots, researchers and developers exploring interpretability within AI systems, market researchers and data analysts leveraging ChatGPT to enhance survey design and extract invaluable insights from consumer data, among others. By actively incorporating user feedback, Chat-GPT continuously enhances the quality of its responses and overall user engagement, thereby elevating its interpretational capabilities. This dynamic feature extends to intelligent survey creation, facilitating accurate data collection, and analyzing unstructured customer input to unveil intricate patterns in consumer attitudes and behaviors. Such versatility proves beneficial to diverse user segments, ranging from individuals seeking genuine and meaningful conversations to researchers and developers delving into AI advancements, technology enthusiasts keen on emerging developments, and content creators harnessing the untapped potential of dialogue generation.

ChatGPT enhances students' writing skills by injecting excitement and curiosity into their writing process. It offers personalized assignments, real-time suggestions, and expert guidance, tailoring prompts to suit each student's unique preferences and skill levels. Leveraging

AI technology, this inclusive approach not only bolsters writing proficiency but also cultivates an engaging and enriching writing experience for various stakeholders, including students, educators, and educational institutions. Moreover, when AI models are transparent and explainable, they lend meaning to the decisions they make, instilling trust in the model and its capabilities [149].

ChatGPT's remarkable linguistic capabilities hold significant implications for the education sector. It transforms student interactions, offers valuable instructional assistance, serves as a user-friendly search tool, and greatly contributes to enhancing writing skills. The educational benefits derived from ChatGPT's capabilities have the potential to reshape how information is accessed and shared, resonating strongly with students, educators, and anyone who values a conversational approach to searching for information.

ChatGPT fosters the cultivation of outstanding writing abilities by prioritizing progress and advancement. It aids digital marketers in crafting impactful and compelling advertising content, thereby enhancing operational effectiveness, while also delivering precise AI assessments and tailored feedback. This functionality caters to a diverse range of users, including writers, content creators, educators, trainers, and professionals in digital marketing. The invaluable support provided by ChatGPT in skill refinement and the crafting of impactful communication stands to benefit all these stakeholders.

ML serves as the backbone for generative AI systems like ChatGPT, enabling them to generate unique content based on existing information. This transformative capability holds the promise of revolutionizing work environments and offering invaluable insights for market research and relationship-building efforts. A diverse array of beneficiaries stand to gain from this innovation, including knowledge workers, researchers, developers, content creators, market analysts, and everyday consumers seeking insightful and imaginative solutions. Leveraging extensive training on human-generated online content and predictive text capabilities, OpenAI's ChatGPT generates AI-powered responses spanning a wide spectrum of topics and writing styles. This versatile tool appeals to a broad audience, including art enthusiasts, poets, creative writers, authors, content creators, scholars, students, and researchers. Through the automation of processes, streamlining of workflows, and enhancement of cybersecurity via virtual assistants and generative AI technologies, ChatGPT enhances the execution of routine office operations. The beneficiaries of this capability encompass office staff and employees who rely on seamless support to accomplish their daily tasks effectively.

3.2 ChatGPT Language Generation Model-Attention Model and RL for LLMs

The ChatGPT language generation model is built on a pre-trained transformer architecture, specifically tailored for natural language processing (NLP) tasks, particularly in the domain of conversational AI. The core components of this generation model comprise:

- Architecture: The architecture of ChatGPT is rooted in transformers, which excel at capturing context from input queries and establishing relationships with prompt sequences. It leverages a self-attention mechanism to focus on different parts of the input sequence, facilitating the generation of responses. This inherent design makes ChatGPT remarkably efficient in producing natural language text.
- Pre-training: In this phase, ChatGPT learns to predict the next word in the output response by training on a large corpus of text data from the internet. This process enables the model to develop a comprehensive understanding of common language usage and context.
- Finetuning: After the initial pre-training phase, the model must undergo fine-tuning on a specific dataset to optimize its performance for that particular domain. This fine-tuning process customizes the model's behavior for specific tasks, such as question-answering and chatbots.
- Conversational ability: It is engineered to engage in conversations akin to human interactions. With the ability to grasp the user's queries, comprehend context across multiple exchanges, and generate responses that are contextually relevant. This makes the system suitable for virtual assistants of any business application.
- Deployment: It finds applications in services demanding human-like responses across multiple queries of the same context, including virtual assistants and customer support, which may be either voice or text-based.

3.2.1 Attention Model

Attention models in transformers play a vital role in capturing dependencies within the input sequence. They calculate the weighted sum of the input query, considering both relevance and position in the sequence, achieved through scaled dot-product attention. This mechanism enables the model to prioritize specific parts of the input data during output generation by assigning varying levels of attention. The three primary attention mechanisms are:

- Self Attention: It identifies dependencies and relationships among tokens in the input query, enabling the model to learn contextual information effectively.

- Encode-Decoder Attention: The encoder processes the input, while the decoder handles the output sequence, aligning the input and output by transferring information between them.
- Multi-Head Attention: It executes the attention mechanism multiple times concurrently. This parallel processing involves utilizing learned parameters from distinct sub-spaces to perform multiple attention operations simultaneously.

The attention score between query Q and key-vector V is defined as follows.

$$Att(Q, K_i) = \frac{Q.K_i}{\sqrt{d_k}} \tag{10}$$

where Q is the query vector derived from the current position of the words, which can be learned by $x_i.Wq$ where $x \in [t_1, t_2,..., t_n]$ denote tokens of the input query and W is the learned value weight matrix. Key vector K is derived from the position of the token in the input sequence, which is learned by $x_i.W_k$. $\sqrt{d_k}$ is the dimensions of the key vector. The attention weight is calculated by the following equation.

$$Q.KAttw(Q, K) = Soft_{max} \frac{Q.K}{\sqrt{d_k}} \tag{11}$$

The weighted sum of the value vector is calculated as follows.

$$Atto(Q, K, V) = \sum_{i}^{n} Att_w(Q.K_i).V_i \tag{12}$$

The value vector V contains the information associated with each position in the input sequence, which is learned by $x_i.W_v$.

3.2.2 Reinforcement Learning in LLMs

Integration of Reinforcement Learning (RL) in LLMs enables the model to perform sequential decision-making and interaction with the system. It uses a special signal to guide the model to make better decisions. The objective function of integrating RL in LLM can be defined as follows.

$$MAX_\theta.E\pi \left[\sum_{\infty}^{t=0} .Y^t.R(s_t, a_t) \right] \tag{13}$$

where MAX_θ defines the optimization that optimize the parameter function θ, and θ are the parameter of LLM. E_π represents the expectations of which behavior or policy a model would use. $\sum_{\infty}^{t=0}$ is summing from time stamp 0 to infinite horizons of RL. Y represents the importance of reward in the future, which ranges from 0 to 1. $R(s_t, a_t)$

represents the reward function that assigns a scaler value to the internal state of the model(s_t) and action of generated text sequence(a_t).

4 Open Issues and Challenges

In this section, we outline the open issues and research challenges related to the integration of ChatGPT into various domains. The details are as follows:

4.1 Open Challenges

Despite ChatGPT's major contributions to scientific advances, it is critical to recognize and address the research challenges that have arisen as a result of its adoption. Table 7 presents an overview of future research options and probable directions. This part investigates these problems and evaluates ChatGPT's potential in the scientific community through emerging developments. The following are common concerns that arise when using ChatGPT for scientific research.

Table 7 indicates the open issues of the ChatGPT. Trust in information generated by AI must be maintained by consistency and precision [150]. Despite ChatGPT's outstanding humanoid writing skills, certain errors and inaccuracies have been found. It is crucial to increase the accuracy and coherence of AI-generated information in order to guarantee the validity of scientific discoveries. Concerns regarding AI-generated unfairness are raised by the usage of enormous amounts of text to train ChatGPT [151]. The AI model may unintentionally reinforce biases found in the training data, thus affecting subsequent research and studies. In order to stop the spread of unfair and biased information, this emphasizes the necessity of addressing biases in the training data.

The pervasive reliance on advanced AI models such as ChatGPT may potentially compromise researchers' autonomy and critical thinking skills. Therefore, it is imperative to strike a balance between leveraging AI technology and preserving researchers' independence and analytical capabilities [152]. While ChatGPT has the capability to generate high-quality language, there is a risk of producing inappropriate or low-quality responses. To ensure consistent production of top-notch content, continuous monitoring, learning, and improvement initiatives are essential [153].

The biases included in the training data can have an impact on ChatGPT's performance. When utilized to produce predictions, biased data can have negative consequences in industries including healthcare, manufacturing, law enforcement, and employment [154, 166]. It is crucial to take into account the volume and variety of data utilized for training AI models in order to reduce this risk. Due to its massive dataset training, ChatGPT frequently has trouble

Table 7 Open issues of the ChatGPT

Future research direction	Description	Application Impact
Consistency and precision [150]	ChatGPT shows impressive human-like writing, but occasional inaccuracies raise concerns for trusting AI-generated data in science. Precision and consistency are key for reliability	Medical Research and Language Translator
AI-generated unfairness [151]	Training ChatGPT on massive text volumes can introduce inherent flaws that the model may unintentionally propagate, potentially impacting future research	Medical Research, Education, and Coding Assistance
Misplaced faithinAI [152]	Excessive reliance on sophisticated AI modelslikeChatGPTcanhamper researchers' autonomy in thinking and problem-solving	Education, Medical Research, and NLP
Quality assurance [153]	Although ChatGPT can produce great language, it can also provide poor-quality or inappropriate responses. To guarantee that ChatGPT constantly produces top-notch content, continuous observation, learning, and development are crucial	Medical Research
Biasinthe data [154]	ChatGPT's outcomes are influenced by the quantity and diversity of training data, and using biased data in fields like healthcare, law enforcement, and employment can lead to detrimental impacts on predictions	Medical Research, Education, and Natural Language Generation
Generalizability [155]	Due to ChatGPT's dependency on huge datasets, its accuracy is frequently hampered, and it has trouble generalising to new data	Medical Research, Virtual Assistance, and Virtual Role Playing in Gaming
Explainability [156]	It can be tough to comprehend the intricate ChatGPT model, making it difficult to understand its decision-making process and spot any errors	Medical Research and Education
Power utilisation [157]	ChatGPT models consume significant computational power due to their information volume, potentially impacting the ecosystem. Enhancing the power efficiency of ChatGPT models offers ample opportunities for improvement	Medical Research, Virtual Assistance, and Virtual Role Playing in Gaming
Immediate feedback [158]	WhileChatGPTcangeneratetext promptly, its response time can be lengthy. Enhancing ChatGPT to be faster and more adaptable would greatly benefit many users	Medical Research and Virtual Role Playing in Gaming
Security issues [159]	When it comes to ChatGPT, it's critical to prevent the emergence of negative content, such as bias and misrepresentation. To solve these issues, it is crucial to create safeguards	Medical Research, Virtual Role Playing in gaming, and NLP
Dataprotectionissues [160]	The creation of appropriate regulations and laws for ethical data processing is necessary to address ChatGPT's access to substantial user information and protect user privacy and security	Medical Research, Virtual Role Playing in gaming, and NLP
Language and culture prejudice [161]	Addressing potential biases in ChatGPT necessitates the use of culturally and linguistically inclusive training datasets and assessment measures to mitigate unsuitable or prejudiced responses	NaturalLanguageProcessing and Language translator
Explanation in models [162]	Enhancinginterpretabilityandtransparency in AI language models like Chat-GPT fosters trust, empowering users to make informed decisions and defend the generated material through transparent decision-making processes and operational information	Medical Research, Virtual Assistance, and Virtual Role Playing in Gaming

Table 7 (continued)

Future research direction	Description	Application Impact
Revising to enhance pro- ficiency in a specific domain 163	To fully leverage the potential of AI lan- guage models like ChatGPT, they need effective adaptation and fine-tuning to spe- cific domains, sectors, and use cases, as they may lack com- prehensive expertise in specialized tasks despite their broad skills and understanding.	Medical Research and Virtual Assistance
Knowledge of contex- tualization 164	Enhancing ChatGPT's ability to maintain consist- ency, understand the broader con- text, and retain meaning in extended con- versations is crucial to overcome ongoing challenges in its performance	Natural Language Processing and Language transla- tor
Correctness in the facts 165	Addressing the challenge of unreliable and inaccurate content generated by AI lan- guage models like ChatGPT is crucial for enhancing their efficiency, accuracy, and utility, enabling advancements in the field.	Education, Medical Research, and NLP

generalizing to new data [155]. In order to strengthen Chat-GPT's capacity to generalize and perform better, new training methods must be created.

Understanding the ChatGPT model's decision-making process and spotting its shortcomings is challenging due to its intricacy. Enhancing explainability is crucial to fostering confidence in AI models and making it possible to see any faults [156]. ChatGPT models use a lot of computational resources, which can be harmful to the environment. In order to lessen the ecological impact of ChatGPT models, there is a need to improve their power efficiency. While ChatGPT may produce text quickly, its responses can occasionally be delayed [157]. Users that want quick and flexible interactions would benefit from ChatGPT's increased speed and adaptability. It is essential to guarantee the security and integrity of content produced by ChatGPT [158]. To stop the formation of detrimental or harmful content, such as intolerance, precautions must be implemented.

Concerns about the privacy and security are brought up by ChatGPT's access to a significant amount of user data. It is crucial to establish the proper procedures and regulations to protect user information and guarantee ethical usage [159]. ChatGPT might have prejudices against particular languages and cultures, which might lead to improper or biased responses. More inclusive training datasets and evaluation criteria are required to solve this problem in order to counteract language and cultural stereotypes [160]. The examination of ChatGPT's security threats reveals several potential weaknesses that warrant careful consideration. Scholars and experts in the field are actively scrutinizing the model's design to identify and address any vulnerabilities. Key areas of concern include susceptibility to adversarial attacks, which could manipulate responses and compromise the accuracy of information provided. Additionally, privacy concerns are being addressed, with a focus on ensuring secure management of user data and preventing unintentional disclosure of private information by Chat-GPT.

Ongoing research aims to bolster the model's defenses against potential attacks exploiting its learning capabilities to extract private user information. Efforts are underway to integrate robust security measures such as anomaly detection, encryption protocols, and reinforcement learning approaches into ChatGPT to enhance its security posture. These endeavors seek to strike a balance between the model's conversational capabilities and the imperative to safeguard user privacy and data integrity. As ChatGPT evolves, ensuring a secure and trustworthy conversational AI experience will necessitate continuous vigilance in identifying and mitigating security threats.

It is essential to increase the outcomes provided by Chat-GPT and other AI language models' interpretability and justifiability. If models become more explicable, transparent in decision-making, and provide information on their working, the trust may be maintained and users can make better decisions [161]. Despite having a wide range of abilities and information, ChatGPT could not have the specialized knowledge required for some activities. To fully realize the potential of AI language models in particular domains, industries, and use cases, effective adaptation, fine-tuning, and its explainability are necessary [162]. Though ChatGPT can produce logical and context-sensitive responses, it might have trouble understanding larger ideas or staying consistent over the course of lengthy chats. The model's capacity to comprehend and hold onto meaning over large information chunks has to be improved and enhanced as per the feedback [163].

It is imperative to guarantee the accuracy of facts generated by ChatGPT and other AI language models, especially

in situations where exact information is essential. A crucial problem that needs to be solved is ensuring the accuracy and compliance of the created material with the supplied data [164]. By addressing these issues, the AI research community may improve the effectiveness, usefulness, and efficiency of language models like ChatGPT, paving the way for the creation of more complex and morally upstanding AI-driven applications in a variety of fields [165].

As indicated in aforementioned discussions, the challenges are directly or indirectly derives from the bias. Hence, bias plays an important factor that needs to be observed for language generation. Bias significantly impacts the fairness, inclusivity, and accuracy of generated content, constituting a pivotal aspect of language production. Language models like ChatGPT are often trained on extensive datasets that inadvertently reflect societal biases, including those related to race, gender, and culture. The model's tendency to replicate biases present in its training data can result in the generation of biased or prejudiced responses. The implications of biased language generation are profound, with the potential to perpetuate injustice, prejudice, and inequitable treatment across various domains such as healthcare, law enforcement, and the workplace. Moreover, biased language production may amplify existing prejudices, propagate negative stereotypes, and further marginalize specific racial or ethnic communities.

Biased content not only perpetuates falsehoods and prejudices but also erodes trust in AI systems. To foster ethical and responsible AI applications, it's imperative to tackle bias in language production. This involves identifying and mitigating biases in training data, promoting inclusivity and diversity in dataset collection, and establishing objective evaluation metrics. Moreover, ongoing efforts to develop algorithms and strategies that mitigate bias in language models and enhance their fairness and inclusivity are crucial. Ultimately, the aim is to create AI systems that deliver more equitable, accurate, and unbiased outcomes.

A complete analysis of societal biases in language production is offered in Table 8, which is motivated by the importance of assuring fairness in language generation. It is important to evaluate developments in bias analysis and reduction since NLG approaches play a role in prejudice perpetuation. The two basic kinds of NLG activities are converting text from one form to another and creating text continuations based on a prompt. An organized summary of various research initiatives concentrating on bias in NLG tasks is presented in Table 8.

4.2 Natural Language Generation: The Evolution and Future Directions

Natural Language Generation (NLG) is a dynamic domain within AI that centers on systems capable of producing

Table 8 Demographic Bias Studies in Various NLG Tasks

Keywords	NLG outcomes	Issues related to the bias and its learning
Gender Identity	Autocomplete	[9, 167–176]
	Dialogue	[177–180]
	Machine Translation	[181–200]
	Re-writing	[201–204]
Occupation, Career, Job, Employment, Profession	Autocomplete	[170];[205]
Race, Nationality	Autocomplete	[167, 169, 170, 174, 206]
	Dialogue	[177]
Religion, Spirituality, Devotion, Religious belief	Autocomplete	[9, 174, 207]
Sexuality	Autocomplete	[167, 208]
	Dialogue	[177]
Other	Autocomplete	[167, 170, 205, 209, 210]
	Dialogue	[211]
	Re-writing	[212, 213]

coherent and human-like text from structured data. Its primary objective is to bridge the gap between raw data and understandable narratives, facilitating the conversion of complex information into readable and relevant text. NLG encompasses various approaches, with two prominent types being Data-to-Text NLG and Text-to-Text NLG. Data-to-Text NLG primarily focuses on transforming structured data, such as tables or databases, into understandable narratives. Conversely, Text-to-Text NLG involves generating natural language text based on existing textual inputs, encompassing tasks such as summarization and translation.

1. Data-to-Text NLG: Data-to-Text NLG stands as a specialized branch within NLG, dedicated to the conversion of structured data, typically presented in tables or databases, into natural language text. Its primary aim is to produce human-readable narratives that effectively convey the information inherent in the structured data. The evolution of Data-to-Text NLG owes much to transformative advancements in technology, particularly the rise of transformer-based models. Transformer architectures, exemplified by models like GPT, have showcased remarkable prowess in comprehending and articulating coherent text. The application spectrum of Data-to-Text NLG spans diverse domains, enriching communication and comprehension of structured information. From financial reporting to weather updates and sports commentaries, Data-to-Text NLG systems find utility in various fields. However, challenges persist in this domain, including the assurance of generating accurate

and contextually relevant text, as well as the handling of ambiguous data.

2. Text-to-Text NLG: Text-to-Text NLG stands at the forefront of advancing language-centric tasks, enabling systems to comprehend and generate coherent text based on existing textual inputs. This approach harnesses the power of advanced deep learning architectures, with transformer-based models like the GPT series emerging as key players. These models leverage self-attention mechanisms to capture intricate contextual dependencies within the input text, thereby facilitating the generation of high-quality outputs. Further enhancements, including fine-tuning on task-specific datasets and the implementation of diverse decoding strategies, contribute to the overall performance of Text-to-Text NLG systems. The impact of Text-to-Text NLG spans across various domains, with notable advancements witnessed in abstractive summarization, paraphrasing, and the generation of diverse linguistic outputs while preserving original meaning. Applications in content creation, information retrieval, communication, and translation have witnessed significant breakthroughs due to Text-to-Text NLG innovations. However, despite its numerous advantages, Text-to-Text NLG encounters challenges related to controllability, domain adaptation, and ethical considerations.

NLG tasks can be broadly categorized into two groups: those that transform text from one form to another and those that generate text continuations based on a given prompt. Tasks such as autocomplete and dialogue creation belong to the latter category, focusing on generating coherent and relevant text in response to a prompt. Autocomplete generation involves producing text directly from language models while adhering to specific constraints. Evaluating biases in these models is crucial for quantifying their impact, particularly in NLG and NLU tasks involving language models.

On the contrary, dialogue generation relies on user inputs and can be tailored to specific domains (such as healthcare or customer service) and functionalities (such as behavior intervention or flight booking), or it can simply involve casual conversation. Any biases present can directly impact user behavior and actions because these dialogue systems interact directly with users. Additionally, when translating text into another language using machine translation, it's essential to maintain the intended meaning. Gender biases in machine translation have garnered significant attention, with studies revealing biases in both academic and commercial systems. Gender associations, such as those with certain professions, may inadvertently arise during translation due to the presence or absence of grammatical gender in different languages. Rewriting tasks involve editing specific words and phrases from the original content to better align with

the desired attribute. Often, these tasks rely on specialized encoder-decoder models.

5 Security and Ethical Considerations

The section delves into the security and ethical considerations surrounding ChatGPT. Through carefully crafted prompts, the GPT model can be manipulated into generating security-related vulnerabilities, which could potentially be exploited by malicious actors. Furthermore, ethical concerns regarding the utilization of GPT are subject to frequent questioning. These considerations are explored in detail below.

5.1 Security Considerations

One of the foremost concerns regarding the potential misuse of ChatGPT revolves around various cybersecurity risks and challenges it poses. These include activities such as fabricating false information, spreading misinformation, or impersonating individuals. The model's remarkable proficiency in generating text that closely mimics human language poses a significant challenge in distinguishing authentic content from fake, thus increasing the likelihood of disinformation and fraudulent activities. Moreover, significant privacy concerns arise due to the extensive and sensitive nature of the data utilized in ChatGPT's training process [214]. The extensive dataset used in training ChatGPT may include private data, which, if compromised, could be exploited to the detriment of users. Therefore, there is an urgent need for robust security measures to safeguard both the model itself and the confidential data upon which it relies.

The risk of ChatGPT-generated outputs being exploited to disseminate false information and deceive individuals is a pressing cybersecurity concern. Such manipulation could empower malicious actors to craft convincing phishing emails or impersonate others, leading to severe repercussions like data breaches or unauthorized access to sensitive information. Consequently, there is a critical need to devise strategies for detecting and thwarting such malicious activities. Particularly, businesses are deeply concerned about ChatGPT's potential to generate persuasive phishing emails capable of deceiving individuals into divulging personal data or clicking on harmful links. Moreover, as ChatGPT and similar AI technologies gain traction, they become prime targets for cybercriminals and fraudsters aiming to exploit them for nefarious purposes. While it's challenging to pinpoint whether particular malware has been directly created using ChatGPT, there have been reports of cybercriminals leveraging the platform to generate code for illicit activities on underground web markets. This underscores the potential for ChatGPT to inadvertently facilitate cybercrime and illicit conduct, prompting concerns regarding its

responsible use and security safeguards. While ChatGPT may offer benefits for security tasks and streamline the work of security specialists, it's imperative to address the cybersecurity implications it presents. This entails implementing stringent security protocols, promoting responsible and ethical usage, and maintaining vigilant oversight over its capabilities through regular updates. These proactive measures are essential for mitigating risks and safeguarding against potential threats. Table 9 summarizes the most significant cybersecurity threats connected with ChatGPT. One significant concern surrounding ChatGPT is its potential to generate private or sensitive information due to its extensive data learning capabilities. This raises the risk of generating unauthorized sensitive data, including the creation of convincing yet fraudulent content like realistic fake conversations and emails, which could be exploited for fraudulent activities or impersonation. Moreover, there's the possibility of generating sensitive information such as medical diagnoses or financial transactions, posing risks to individuals' well-being and identity security. To mitigate these potential risks, it's crucial to enact stringent privacy regulations and maintain ongoing oversight.

Another significant risk associated with ChatGPT is its potential for tailored fraud and deceptive tactics. It could be used to craft messages that appear to be from reputable sources, such as banks or trusted acquaintances, with the intent of gathering sensitive information or financial gain. Moreover, there's the concern that ChatGPT may generate inflammatory or inaccurate content, leading to the spread of misinformation and potential incitement of violence. From a technical perspective, there's the issue of the model's capacity to retain and reproduce personal information from the data it has learned, as evidenced in past instances involving models like GPT-3 and GPT-2. These instances underscore the importance of exercising caution when deploying such

models with sensitive data to mitigate privacy breaches and data security risks, emphasizing the necessity for meticulous data handling and ongoing model supervision.

Analyzing the relationship between security issues and ChatGPT underscores the critical importance of real-time monitoring in safeguarding the integrity of the model. Real-time monitoring encompasses a suite of surveillance mechanisms meticulously designed to swiftly detect and address potential security risks that could compromise ChatGPT's functionality or user privacy. These surveillance methods are finely tuned to identify various anomalies, including potential privacy breaches endangering sensitive user data and adversarial attacks aimed at manipulating the model's responses. A detailed examination of ChatGPT's real-time monitoring mechanisms reveals their capacity to detect anomalies such as inconsistencies in model output, deviations from expected user interactions, or irregularities in input patterns. Moreover, a comprehensive overview of the proactive security measures integrated into the system outlines the corresponding response tactics for each identified anomaly. Real-time monitoring is indispensable for promptly identifying potential security breaches, thereby facilitating risk mitigation and ensuring swift and effective response measures. By employing dynamic surveillance methods, ChatGPT demonstrates its commitment to adhering to stringent security standards and adapting to evolving threat landscapes. Emphasizing the value of real-time monitoring not only underscores the model's resilience against diverse security threats but also enhances user confidence by providing reassurance regarding its robust security posture. Providing users and stakeholders with a comprehensive understanding of the real-time monitoring protocols fosters trust in ChatGPT's security mechanisms and bolsters confidence in its ability to navigate complex security environments effectively. However, certain security issues with ChatGPT can

Table 9 ChatGPT security risks

Risks associated with the Cybersecurity	Description
Unsecured information	The data that ChatGPT uses might have sensitive information, and if someone gains access to the model or the data, they could misuse it
Malicious takeovers	Cybercriminals might try to take control of ChatGPT and use it for bad things, like spreading false information or causing harm
Malicious software contamination	People could utilize ChatGPT to construct appealing fake messages which deceive people into downloading or running malicious malware
Inappropriate access	If the security checks or safeguards aren't strong enough, unauthorized people might get into ChatGPT and use it inappropriately
Brute force attacks	Cybercriminals might try to figure out passwords or access controls for ChatGPT, which could let them get in without permission
Availability	ChatGPT's ability to function properly could be disrupted by attacks like DDoS or spam attacks
Overload of information	ChatGPT might have trouble dealing with a lot of information, which could make it work slower or make mistakes

be mitigated. Table 10 provides a quick overview of the technological actions we may take to reduce the possibility of ChatGPT producing personal data when fine-tuned with this sort of information. It's vital to remember that the effectiveness of these processes varies based on the situation and data employed. A combination of technological, organizational, and legal efforts is the most effective strategy to reduce risks.

5.2 Discussion of Ethical and Societal Implications

LLMs have affected the general fairness and ethical considerations associated with these models by providing rise to a variety of ethical concerns around employment, misinformation, privacy, bias, power dynamics, transparency, and intellectual property [215]. The primary concerns are mentioned in the Table 11

As a language model within the realm of AI, ChatGPT is engineered with the primary objective of leveraging patterns and probabilities derived from extensive data analysis to comprehend and generate responses to user inputs. However, the utilization of this technology may inevitably give rise to significant ethical considerations [216]. In order to clarify this, Table 13 mentions few examples of these issues.

Table 10 Technical safeguards to reduce the possibility of ChatGPT creating personal information

Technical Standards	Explanation
Differential Privacy	To obscure an individual's identify, add noise to the data
Multiparty computation security	Do calculations with private data without showing what the data is
Federated Learning	Train models to learn from data on different devices or in different organizations without giving away the actual data
Encryption	Keep data safe from unauthorized access
Detection of the anomaly	Identify and mark whenever the model creates private or sensitive information
Access control	Make sure that only the right people who are allowed to do so can use the model and see what it produces
Monitoring and review on a regular basis	Keep a close eye on how well the model is doing its job and what it's producing. Look out for any privacy problems, and if you find any, make sure to fix them

Table 11 Ethical Concerns of the LLMs

Ethical Concerns	Description
Privacy	LLMs need vast data for training, including sensitive information, which, if mishandled, can harm individuals' privacy
Bias	LLMs can amplify biases from their training data, perpetuating and worsening societal biases, such as racial and gender discrimination
Transparency	LLMs are often seen as"black boxes," making it hard to understand their predictions and address biases or ethical issues
Power	LLMs can shape public discourse and decision-making, granting significant power and responsibility to their creators and users, with potential risks of misuse for opinion manipulation or dissent suppression
Intellectual property	A debate on intellectual property rights revolves around LLMs, focusing on data and model ownership, with potential far-reaching implications for intellectual property law
Employment	Large language models, in their ability to automate tasks once carried out by humans, raise concerns about potential job displacement and the necessity for workforce retraining
Misinformation	LLMs can create false information, swiftly spreading through online platforms, leading to harmful consequences like election interference or incitement of violence

Table 12 Potential ethical issues of the ChatGPT

Potential Issues	Description
Bias	ChatGPT can inherit and perpetuate biases from its training data, reflecting and sustaining those biases
Misinformation	ChatGPT may respond with erroneous or misleading information, especially if it isn't trained on trustworthy sources. When users rely on ChatGPT for direction or counsel, this could be dangerous
Privacy	User information, maybe including personal information, may be gathered and stored by ChatGPT for unforeseen uses such targeted advertising or monitoring
Responsibility	Because ChatGPT is morally impotent, it cannot be held responsible for its deeds. However, it is the responsibility of those who developed and implemented ChatGPT to guarantee its moral application and guard against user harm

It is crucial to identify and acknowledge these ethical issues in order to demand aggressive steps to address them and ensure that ChatGPT is used responsibly and ethically. LLMs, like ChatGPT, can use a variety of approaches and tactics to start a path toward maintaining equity in their responses. The Table 12 follows few tactics that ChatGPT can use to actively promote equity in the responses it generates.

5.3 ChatGPT Applicability and Penetration Towards Future NLP Model Designs

As ChatGPT underwent extensive training on a large dataset of online text (approximately 570 GB in total), it signifies a significant milestone in the advancement of Natural Language Processing. The model has acquired the ability to comprehend intricate linguistic patterns and probabilities through this rigorous training, enabling it to generate responses that are contextually logical and appropriate. With a low perplexity score of approximately 35, indicating its accuracy in predicting the next word in a sentence, ChatGPT demonstrates a notable improvement over previous models and reflects a deeper understanding of language context.

This enhanced capability, offering more precise and context-aware language generation, marks a notable advancement in NLP technology. ChatGPT has made a substantial impact on Conversational AI, leveraging the advantages of NLP. For instance, GPT-4 demonstrates improved context comprehension, enabling chatbots to provide more contextually relevant responses. It is worth noting that the model's training data is derived from a diverse set of internet text sources [217] and the same has been depicted in Table 14.

5.3.1 Influence of ChatGPT on NLP Models

NLP models such as BERT-2 [218] have been inspired directly by ChatGPT's effectiveness in generating coherent and contextually relevant text responses. While ChatGPT excels in generating human-like conversation, models like as BERT-2 concentrate on improving specific parts of NLP tasks such as information retrieval. This effect is most noticeable in BERT-2's pre-training method, which lays a high emphasis on enhancing the model's comprehension of language in order for it to succeed at information retrieval tasks. As an example influenced by ChatGPT, BERT-2 is developed with a strong emphasis on pre-training. The initial step in which a language model learns from a large corpus of text is referred to as pre-training. What distinguishes BERT-2 is its commitment to improving the pre-training phase, ensuring that the model fully understands linguistic nuances and context.

By understanding the nuances of language usage, this rigorous pre-training prepares BERT-2 to succeed in different NLP tasks, particularly information retrieval. The BERT-2 approach systematically selects training data from diverse sources to achieve its pre-training objectives. This diversity is crucial for exposing the model to a wide array of linguistic styles, domains, and usage contexts. The training data for BERT-2 includes extensive corpora sourced from various mediums such as internet text sources, books, Wikipedia articles, and news stories. These diverse data sources play a crucial role in enhancing the model's understanding of language patterns, enabling it to proficiently comprehend and generate text across a broad spectrum of themes and genres. BERT-2's extensive corpora encompass text sourced from the internet, rendering it a vast and comprehensive repository. This corpus comprises user-generated content, news articles, academic

Table 13 Imposing fairness in the ChatGPT responses

Addressing Bias and Promoting Fairness in Chat-GPT Responses	Description
Data Diversity and Inclusion in Training	ChatGPT reduces bias by diversifying its training data, aligning it with a more diverse range of human communication patterns and usage scenarios
Counterfactual Data Augmentation	By creating instances that oppose training data biases, counterfactual data augmentation helps ML models become less biased and makes it easier for ChatGPT to identify and lessen these biases
Debiasing Methodologies	ChatGPT can decrease biases by employing debiasing strategies, which entail changing training data or the model itself to reduce response bias. These methods range from simple data manipulation to advanced pattern detecting systems
Constant Monitoring and Assessment	By routinely assessing and analyzing its responses for bias and obtaining user feedback to correct flaws, ChatGPT can uphold fairness
Instructions for Human Editors	Reviewers are advised by OpenAI to steer clear of prejudice, such as hate speech and discriminatory model replies
Feedback from Users and Community Participation	To encourage continuous changes in the behavior and performance of AI models, such as ChatGPT, OpenAI invites user feedback on problematic outputs

papers, and various other textual forms. Additionally, the inclusion of books and Wikipedia articles enriches the model with domain-specific knowledge and terminology, thereby augmenting its linguistic capabilities. BERT-2 is specifically engineered to address NLP tasks that necessitate domain-specific information and context, leveraging the richness of these diverse sources. In essence, BERT-2's focus on pre-training and meticulous curation of training data reflects the influence of ChatGPT's advancements in NLP. These methodologies empower BERT-2 and similar models to excel in their respective NLP domains, facilitating nuanced language comprehension and enhanced performance, particularly in tasks such as information retrieval where contextually rich understanding is paramount. Few examples for the same has been depicted in Table 15. The challenges and ethical Considerations in ChatGPT and future NLP Models are shown in Table 16. These difficulties and ethical considerations are critical. To mitigate bias, for example, diversified training data and regular bias checks are required. Encryption and access restrictions protect data privacy, while content monitoring and clear standards ensure ethical usage [219]. Chat-GPT and similar NLP models present promising opportunities, but their responsible and meaningful utilization requires

Table 14 LLMs influenced by ChatGPT evolution

LLM	AI Model	Description	Application	Limitations& Future Scope
GPT-4	OpenAI	Multimodal large language model with improved reliability and creativity	Conversational AI, creative writing, image input processing	Addressingbiases, enhancing efficiency, expanding multimodal capabilities
PaLM-2	Google	Multitasking language model excelling across languages	Machine translation, multitask learning	Improving data diversity, extending to more languages and tasks
Claude	Anthropic	Focuses on userligned language understanding for per- sonalized interactions	Personalized assistants, content moderation	Enhancingpersonalization,ensuring ethical AI practices
LLaMA	Facebook AI	Efficient language modeling aimed at sustainability	Knowledge extraction, summarization	Reducingcomputational costs, ensuring data privacy
BERT- 2	Google	Enhanced context understand ing for better information retrieval	Information retrieval, sentiment analysis	Deepening contextual understanding, adapting across domains
T5- Next	Google	Advanced in contextual text summarization	Content summarization, dataannotation	Improving summarization accuracy, devel oping advanced techniques
Jais 30B, Core42	Arabic LLMs	Arabic language applications, cultural content generation, regional language understanding	Focus on a singlelanguage may-limit broader applicability	Expansion to include moreregional-languages and dialects
Embodied AI	Apple	EmbodiedAIindevices, context-aware computing, user interaction	Integration withphysicaldevices, contextcomprehension	Enhancingdevice interaction, expanding embodied AI applications

Table 15 Information retrieval based LLMs

Model Name	Key Features	Application	Training Data Uti- lized
GPT-4	Improved context understanding	Chatbots, Content generation	Diverse internet text data
BERT-2	Enhanced pre-training	Information retrieval	Large-scale corpora and books
T5-Next	Contextual text sum- marization	Contentsummariza- tion	Wikipedia, news articles, books

Table 16 Challenges and Ethical Considerations in ChatGPT and Future NLP Models

Challenge/Ethical Concern	Description	Mitigation Strategies
Bias Mitigation	Ensuring reduced biases in responses and training data	Diverse and inclusive training data, bias audits
Data Privacy	Safeguarding user data and maintaining privacy during training	Encryption, data anonymization, and strict access controls
Ethical Usage	Preventing misuse for harmful purposes, such as misinformation	Content moderation, ethical guidelines, and user reporting mechanisms

collaboration among researchers, developers, and ethical stakeholders. Continuous monitoring of advancements in model design, training data sources, and ethical standards is essential to ensure responsible usage. Chat-GPT's significance in advancing NLP is underscored by its extensive training data and improved contextual understanding, serving as inspiration for subsequent NLP models tailored to specific purposes. However, addressing ethical considerations and challenges is imperative for the future responsible and ethical use of these models [220].

6 Eco-GPT: A Cost-effective and Scalable LLM Design

In the rapidly evolving landscape of information and technology, Language Models (LLMs) have emerged as prominent tools in both research and practical applications. These models have surpassed their original purpose of understanding natural language and generating text. Their capacity to analyze extensive textual data and produce human-like responses has revolutionized our interactions with machines and our access to information. These models find applications across diverse domains, including language transformation, medical research, and scientific discovery. A notable trend is the creation of human-like text through LLMs, often seen in chatbots. Users typically subscribe to interact with such models, with popular examples being ChatGPT and its various versions developed by OpenAI. In this section, we analyze the different pricing structures and strategies users can employ to reduce the inference cost while using LLM. The pricing structure can vary from specific use case and the providers.

- Subscription plane: Providers provide monthly and annual subscription fees to access LLM using API. These plans include a certain number of requests per day or tokens in the request.
- Pay as you go: Providers offer a pay-as-you-go strategy, where the bill will be generated based on actual usage. These flexible usage works well when you are in varying workloads.
- Number of tokens: The Pricing model is based on the number of tokens in the request query that is processed by the model. You will be charged for both input and output tokens, but some complex queries can increase the cost.
- Request-based pricing: Provider will charge you based on the number of API calls you made and billed accordingly.
- Geographical pricing: Prices can vary by region, the geographic location from which you access the service may impact your costs of using LLM.

- Extra add-on: Plugins that are developed by the third party are integrated to offer tons of different features such as instaCart, FiscalNote, Speak openTable, and others.

In Large Language Models (LLMs), estimating query costs involves a multifaceted approach. The primary factor in computing the cost is the length of tokens, which measures both input and output text. Additionally, the frequency of API calls is a crucial factor; the more calls made, the higher the incurred cost. However, utilizing batch API calls, where multiple small queries are combined into a single batch, can potentially lower the cost estimation.

Estimating query costs in Large Language Models (LLMs) involves considering various factors such as the number of tokens used, pricing per token, data transfer costs, batching practices, and the inclusion of special tokens. These parameters collectively determine the overall cost of interacting with an LLM. The cost of querying such a model API mainly comprises three components: (1) the length of the input query, (2) the length of the output generated, and (3) a fixed cost for each query. In our analysis, we examined six different commercial LLMs, including ChatGPT, GPT-3, and GPT-4 from OpenAI, as well as J1-Grande, GPT-J, and X-large from AI21, Textsynth, and CoHere, respectively. To mitigate costs, three strategies are discussed below:

- Prompt Adaptation: It is a process of customizing the prompt i.e. instructions given to LLM to get the desired response. The main objective of the prompt adaptation is to make your query more accurate and specific by defining the context. Optimizing the prompt can lead to obtaining responses in fewer queries, thereby reducing computational costs and improving response time. This refinement facilitates obtaining responses with fewer iterations, reducing the total number of tokens in the output. By obtaining specific responses to our queries, we can write more precise and focused prompts, minimizing irrelevant responses that are later filtered out. Prompt adaptation in automated systems leads to the generation of accurate and less irrelevant data, thereby reducing the need for human interventions and lowering the overall operational costs.
- LLM Approximation: LLM models have high computational overhead due to size and complexity. LLM approximation aims to offer an approximate solution to queries, reducing computational costs. This is accomplished by developing smaller models derived from fully complex LLMs. Although these models require less computation, they retain the capability to generate correct and approximate responses. Reducing model size can be achieved by pruning less important parameters, thereby decreasing computation requirements. Knowledge distillation offers another approach, shrinking model size while retaining

the ability to generate similar results by training with the same data. Converting model weights from floating point to fixed point integers can also decrease computation while maintaining acceptable performance levels. Caching frequently queried data reduces the need for recomputation, saving both time and computational resources. Additionally, augmenting hardware resources, such as incorporating graphics processing units, enhances inference speed and reduces overall costs. However, it's important to note that LLM approximation involves a trade-off between performance and computational cost, as reducing computation may result in some loss of accuracy.

- LLM Cascade: It entails organizing multiple models in a hierarchical structure according to their size and complexity. This hierarchy ranges from simpler, less compute-intensive models at lower levels to more complex and capable models at higher levels. As such, straightforward queries are directed to lower levels, requiring less computation, while intricate queries are forwarded to higher levels. Employing lower-level models for simple queries doesn't demand the full computational power of the LLM model, thereby reducing resource consumption. This tiered model can be tailored to organizations' needs, enabling them to flexibly adjust the number of models at each layer. The LLM cascade improves response time by directing simpler queries to lower levels, thereby reducing the time needed to generate an acceptable response and saving costs. If the lower level fails to produce an acceptable response to a query, a fallback mechanism is employed, rerouting the query to a higher-level model, ensuring that queries are adequately addressed.

6.1 The LLM Cascade Architecture

The "LLM Cascade Architecture," an advanced natural language processing (NLP) method, orchestrates a series of language models (LLMs) in a layered fashion. To grasp its inner workings, it's crucial to understand the functions of each layer, the gradual enhancement in language understanding, and its overall significance to the topic at hand. At the foundational level sits a core language model like GPT-J, designed to capture general language patterns, grammar, and semantics. Acting as the initial filter for incoming queries, its output is evaluated against a predetermined threshold, often set at 0.95 for models like Eco-GPT. If the score exceeds this cutoff, a precise response is generated promptly. However, if the score falls below the threshold, the architecture triggers the next layer, J1-G. This secondary layer enriches contextual awareness and domain-specific knowledge to better understand the input question. Another threshold, typically around 0.35, determines whether J1-G's response meets the criteria or requires further refinement. If neither GPT-J nor J1-G provides a satisfactory

response, the architecture proceeds to the third layer, GPT-4, the pinnacle of complexity. Employing increasingly sophisticated models based on the input's complexity, this sequential cascade structure ensures a dynamic and adaptive approach.

The significance of the LLM Cascade Architecture lies in its ability to address the limitations of relying solely on a single, monolithic language model. By employing a cascade of models, the system can effectively handle a broad spectrum of queries, ranging from simple to complex. This adaptability proves crucial in real-world conversational scenarios, where user inputs exhibit significant variability. The architecture's dynamic nature, capable of adjusting complexity in real-time, enhances its effectiveness and efficiency in delivering contextually appropriate responses. Thus, the deliberate stacking of language models within the LLM Cascade Architecture serves to elevate the level of complexity and flexibility in natural language processing. Its sequential design and threshold-based decision-making ensure a sophisticated and contextually aware process for generating answers, ultimately enhancing language model performance across diverse applications.

Figure 6 presents the architectural diagram of LLM cascade where Qk is the input query that needs to be processed. The reliability score rs n of model Mn is checked against the threshold Th and based on complexity and resource availability, Qk is forwarded to the appropriate Mn i.e. simpler Qk is routed to M1 and complex Qk is forwarded to a higher level that generates response equal to the R. The concept of LLM cascade allocates the queries to different models based on the level of complexity and resource requirement. The allocation can be represented as follows.

$$L_i \leftarrow \text{allocate}(Q_k, M_1, M_2, M_3...M_n) \tag{14}$$

where L_i denote the cascading level, Q_k denotes the requested input query and M_n are the model available at that layer, M_1 is the smallest and M_n is the largest model. To monitor the utilization of resources at each level, organizations need to manage the model at each level.

$$U(L_i) \leftarrow \text{Monitor}(L_i) \tag{15}$$

where $U(L_i)$ denotes allocated computation resources at level L_i and M onotor(L_i) monitors the resource utilization at level L_i the function dynamically adjust the resource to
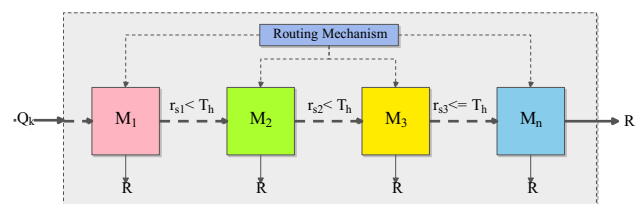

Fig. 6 LLM Cascade Architecture

lower the cost and utilization of model M_n. At each level, it is to be decided whether to route the query to other level based on the complexity.

$$M(Q_k) \leftarrow route(Q_k, M_1, M_2, M_3...M_n) \qquad (16)$$

where M (Q_k) denote the new model on which the query is routed route() function routing the query based on the characteristics and resource requirement. The overall cost can be evaluated based on the Ru resource utilization, R_t response time, reliability of the query score of Qk, and response of the model API M_{API}.

$$cost \leftarrow Cal(R_u, R_t, Q_k, M_{API}) \qquad (17)$$

The main cost of LLM cascade depends on LLM router and scoring function. The scoring function is defined as:

$$G : QkXMAPI \rightarrow [0, 1] \qquad (18)$$

This function computes the reliability score between 0 and 1, the function can be obtained by training the regression model that learn itself by the correct answer and the generated response.

6.2 Prompt Engineering

Language models like LLM are trained on vast corpora of textual data, enabling them to generate text that closely resembles human language. Prompt engineering involves crafting precise instructions to elicit desired responses from these models. It serves as a strategic approach to harnessing the full potential of LLMs and ensuring they grasp the task at hand. The fundamental aspects of prompt engineering include:

- Job Specification: Well-designed promts are essential when Job or task is specific. The prompts need to be given with the specific task, such as a"concluding remark" that summarizes the output text.
- Information context: To generate meaningful responses, the context of the prompt needs to be set that depends on the complexity of the input query i.e. you need a response as a "business analyst".
- Fairness: Ethical consideration while generating prompt is very important. Practitioners should know the biases of the response and are advised to design the prompts that handle the bias.
- Safety and control: To ensure appropriate and safe responses, prompts need to be engineered in such a way that it follows the constraints and behavior of the LLM.
- Adaptation: Based on the domain on which these models are applied needs to adapt specific terminologies of that

domain. This helps the model to respond in the context of the input query.
- Iteration: The practitioner needs to experiment with different input prompts to optimize the output of the query. This is an iterative process to achieve the desired response from the LLM.
- Feedback: Evaluation of the feedback is important from the user that helps fine-tune the model and improve the reliability of the response.

6.3 Performance Evaluation

In this section, we discuss the details of the results obtained from the experiments. For the implementation, 6 different commercial LLM are identified.
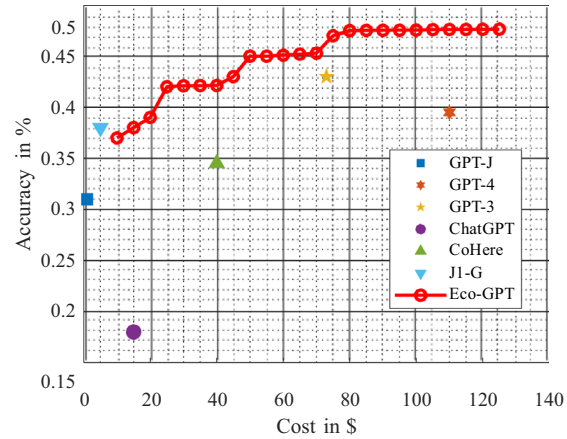
6.3.1 Experimental Setup

We have analyzed 6 different commercial LLM from OpenAI, CoHere, TextSynth and AI21. We have considered ChatGPT, GPT-3 and GPT-4 LLM from OpenAI [221], J1-Grande, GPT-J and X-large from AI21 [222], Textsynth [223], CoHere [224], respectively. Eco-GPT has been implemented on top of these API and evaluated on CaseHold [225] and CoQA [226] dataset. CaseHold dataset consists of 53,000 questions to check the holding of the legal decisions i.e. goal is to check whether a given verdict is rejecting the previous legal case, where CoQA is a conversational question-answering challenge dataset consisting of 127,000 questions with answers collected from different conversations of different domain. For analysis, we kept the cascade length as 3 and dataset is randomly split into training and test set for evaluation [103].

6.4 Performance and Cost

The three levels of cascade consist of GPT-J, J1-G and GPT-4. The Eco-GPT invokes GPT-J, J1-G and GPT-4 sequentially one by one. For any input query, first, it gets the answer from GPT-J, and the output response will be accepted only when a certain threshold value is cleared, for this analysis, we considered 0.95. if the generated score of the response is less than the threshold, J1-G is invoked; if the response score is greater than 0.35, the answer is accepted otherwise, GPT-4 will be invoked to get the final response of the input query. In over observation, this approach performs better than GPT-4 alone in several cases. Figure 7a represents the trade-off between performance and cost, and it is clear from the results obtained that is approximately ≈73% cost saving in CaseHold dataset similarly Fig. 7b represents the trade-off between performance and cost on CoQA dataset and we observed that there is approximately ≈60% cost saving.
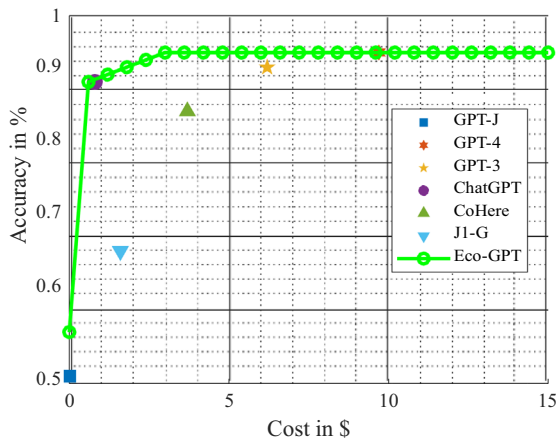
(a) Performance on CaseHold database (b) Performance on CoQA database

Fig. 7 Performance analysis of Eco-GPT on two different database

7 Discussions and Summarized Findings

7.1 Key Findings

A pivotal discovery from the survey reveals that the system adeptly produces responses akin to those crafted by humans across a diverse spectrum of inquiries and prompts. This proficiency significantly heightens user involvement and overall satisfaction levels.

An additional significant discovery underscores ChatGPT's capacity to acquire and adjust to novel scenarios and contexts, thereby enhancing its ability to furnish nuanced and refined responses as time progresses. Research indicates that ChatGPT demonstrates competence in delivering precise and beneficial answers to intricate inquiries spanning various domains such as medicine, law, and finance. Such findings hold considerable significance for these respective industries.

Yet another pivotal discovery underscores the system's adeptness in crafting responses customized to individual users, drawing from their prior interactions with the system. This tailored approach fosters more personalized and efficacious communication channels. Additionally, research has unveiled ChatGPT's capability to discern and interpret idiomatic expressions and figurative language—a fundamental aspect of human communication that poses challenges for AI systems. Studies conducted on ChatGPT underscore the critical significance of continual development and enhancement of NLP systems. This ensures their continued relevance and utility in an ever-evolving technological environment.

Another significant discovery is that the evolution of ChatGPT has been facilitated by advancements in machine learning (ML) and deep learning (DL) techniques.

These advances have empowered researchers to devise progressively sophisticated language processing algorithms. Additionally, research has demonstrated that ChatGPT can produce responses that are not only more engaging but also more interactive compared to those generated by alternative AI systems. This capability has the potential to enhance user adoption and satisfaction levels.

Finally, an important discovery from research on ChatGPT is its ability to generate contextually relevant responses, considering the broader context of the conversation or query. This capability leads to more effective communication, as the system can tailor its responses to fit seamlessly within the ongoing dialogue or discussion.

Other studies emphasize the significance of assessing the performance of NLP systems like ChatGPT across various metrics, including accuracy, response time, and user satisfaction, to ensure they effectively meet user requirements. Additionally, research on ChatGPT underscores the challenges associated with designing and implementing user interfaces for AI systems, particularly in conveying complex information to users in a clear and comprehensible manner. Another significant discovery is that the advancement of ChatGPT has been facilitated by the accessibility of extensive datasets containing natural language data. These datasets have empowered researchers to train and enhance the system's language processing abilities effectively. Additionally, studies have demonstrated that ChatGPT possesses the capability to discern and interpret sentiment and emotion expressed in human language—a critical facet of successful communication. Such capabilities hold promise for various applications in domains like customer service and marketing. A significant discovery from research on ChatGPT is its ability to generate responses that are contextually appropriate,

considering the social and cultural norms inherent in the conversation or query. This capability leads to more effective and respectful communication, acknowledging and adapting to the nuances of various contexts.

Other studies emphasize the crucial role of continuous research and development in the field of NLP to ensure the ongoing improvement and evolution of systems like ChatGPT. This dedication results in enhanced performance and user experiences over time. Furthermore, research on ChatGPT underscores the significance of prioritizing user privacy and data security, especially concerning the application of AI in sensitive domains like healthcare and finance. A key revelation from investigations into ChatGPT is its capacity to generate responses that surpass the accuracy and helpfulness of those produced by traditional keyword-based search engines. This enhanced capability holds the potential to significantly improve information retrieval processes, offering users more effective access to relevant information [227]. Other studies have highlighted the importance of user feedback in refining and improving AI systems such as ChatGPT, in order to ensure that they are meeting the needs and preferences of the clients.

7.2 Future Research and Development

As research and development in NLP continue to evolve, the future of ChatGPT would be driven by ongoing improvements in language understanding capabilities. One key area of focus for the future development of ChatGPT will be the ability to interpret complex queries and requests more accurately. The ongoing use of DL techniques will be an essential component of the future research and development of ChatGPT, allowing for more nuanced and sophisticated language processing capabilities. Given the computational demands of Large Language Models (LLMs), there is a significant emphasis on improving efficiency and scalability. As these models become more complex, optimizing their training processes becomes paramount. Techniques such as model distillation, mixed precision training, and asynchronous gradient updates are being explored to enhance training efficiency, reduce computational costs, and minimize energy consumption. In various application domains, LLMs find utility in tasks such as machine translation, sentiment analysis, chatbots, text summarization, and serving as natural language interfaces for databases. However, it's imperative to address the challenges posed by misinformation and malicious use, particularly considering LLMs' ability to generate realistic and coherent text. Strategies to mitigate these risks include implementing robust content authentication mechanisms, promoting digital literacy, and establishing ethical guidelines for AI-generated content. The potential avenues for research and development are as follows.

- Zero-shot learning: It is an emerging paradigm that enables LLMs to perform tasks for which the model is never explicitly trained. The current study explores the methodologies, difficulties and use of zero-shot learning in LLM, emphasizing the potential of improving the understanding of natural language and the acquisition of knowledge across several modalities. Potential future prospects in the field of AI encompass the vital role of resolving ethical issues, enhancing the efficiency of metalearning techniques, and fostering effective collaboration between humans and AI entities. These endeavors are expected to facilitate the development of AI systems that are more versatile and adaptive in nature.

- Multimodal LLM: It is a groundbreaking frontier in the field of AI research and development. These models seamlessly integrate text and visual information, allowing for a comprehensive understanding of content in a unified approach. They hold the potential to bring profound changes in various domains, spanning from computer vision to NLP, by enabling robust cross-modal interactions. Ongoing and future research in Multimodal LLMs will be focused on enhancing their capacity to extract insightful knowledge from diverse data sources, thereby rendering them indispensable for an array of applications, including content generation, autonomous systems, and tailored user experiences. As these models continue to advance, they are set to redefine the limits of AI's capabilities.

- Few-shot learning: It involve training LLM to perform task with minimum exam-ple. It entails the training of LLMs to swiftly comprehend and generalize from exceedingly scarce examples, enabling them to adapt to a diverse array of tasks with minimal data. By leveraging methodologies such as meta-learning, attention mechanisms, and fine-tuning, Large Language Models (LLMs) can assimilate new knowledge and perform a multitude of tasks with minimal examples. This advancement effectively closes the gap between pre-trained models and task-specific applications, providing a remedy for the issue of data scarcity. This research domain holds the promise of enabling LLMs to thrive in specialized and niche domains characterized by severely limited training data.

- Controllable generation: It is a dynamic research frontier focused on granting users precise command over the output of these models. This involves the modulation of various aspects, including style, tone, and content, with the ultimate goal of enhancing LLM versatility and adaptability to specific applications. This innovative research provides users with a tool for customizing generated outputs according to their specific needs and at the same time, preserving the exceptional language generation capabilities inherent to LLMs. As these techniques

continue to advance, they hold the potential to revolutionize a multitude of industries, from content creation to customer service, through the facilitation of fine-grained control over the generated text.

- Multilingual LLMs: It serves as a critical component in the field of language under-standing and generation. These models play a vital role in changing linguistic barriers and facilitating effective cross-cultural communication, which makes them important in international diplomacy and global business. LLMs are instrumental in the preservation and promotion of endangered languages. Its utility extends to machine translation, sentiment analysis, and content localization, revolutionizing the accessibility and engagement with digital content on a worldwide scale. Ongoing research and development in this sphere promise to enhance global connectivity and enrich cultural exchanges even further.

- Specialized domain adaptation: We can fine-tune GPT for specific domains such as healthcare, legal, and technical support to enhance accuracy and relevance within those domains. Transfer learning enables GPT to leverage knowledge from one domain to enrich its understanding in another. Furthermore, maintaining longer conversations fosters greater coherence in responses. This can be achieved by implementing a memory augmentation mechanism to extend the model's capacity for retaining context over extended dialogues.

- Bias Mitigation and Explainability: To foster fair and inclusive interactions with diverse users, it's crucial to develop strategies for identifying and mitigating bias in training data. Additionally, researchers are investigating methods for providing explanations alongside each response to offer insights into user interactions, thereby enhancing understanding and trust.

- Ethics and personalization: GPT models must be trained to generate outputs responsibly, integrating mechanisms to prevent and detect harmful or malicious content. Researchers should focus on adopting response strategies tailored to users' preferences, history, and context. This entails developing techniques for constructing and maintaining models that facilitate personalized interactions.

- Privacy: A diversified strategy is pivotal for the future direction of research aimed at addressing privacy concerns and ensuring secure interactions for AI models like ChatGPT. A critical avenue involves exploring and enhancing state-of-the-art privacy-preserving techniques. Researchers are actively investigating secure multiparty computing, encryption protocols, and differential privacy methods to bolster ChatGPT's privacy capabilities and empower it to offer supportive conversational assistance while safeguarding user data. Simultaneously, efforts are underway to enhance the interpretability and transparency of ChatGPT's decision-making processes. By providing users with a clearer understanding of how the AI model manages information, researchers aim to strengthen the commitment to user confidentiality and foster a sense of control and confidence. Ongoing research is dedicated to fortifying the security measures embedded within the architecture of chat-based AI models, alongside privacy-preserving safeguards. Techniques such as adversarial training, anomaly detection, and reinforcement learning for secure decision-making are under close scrutiny to enhance ChatGPT's resilience against potential adversarial attacks. Federated learning remains a focal point for research as it enables models to learn from user interactions in a decentralized manner, ensuring that private data remains on users' devices and mitigating the risk of data breaches or unauthorized access. In conclusion, research efforts surrounding ChatGPT are aimed at cultivating a robust and secure conversational AI experience that aligns with evolving privacy standards.

8 Final Thoughts

ChatGPT is an outstanding example of how far AI technology has progressed in recent years. Its ability to comprehend and synthesise human language has made it a valuable tool for a variety of applications. ChatGPT is poised to revolutionize the way we engage with technology and each other, extending its impact across various domains including chatbots, customer support systems, and educational and research applications. One of its standout features is its ability to generate coherent and contextually relevant responses, made possible by its sophisticated algorithms. These algorithms enable ChatGPT to ingest extensive datasets and connect seemingly disparate pieces of information. As ChatGPT and similar language models progress and enhance their learning mechanisms, they are anticipated to tackle more complex tasks with greater proficiency.

However, despite the promising advancements, concerns arise regarding the potential drawbacks associated with adopting AI language models such as ChatGPT. Chief among these concerns is the issue of bias within the training data used for these models. Should the data contain any form of distortion, such as societal preconceptions or stereotypes, there's a risk that the AI model will learn and perpetuate these biases. To mitigate this risk, developers must take meticulous care to ensure that the training data is diverse, representative, and free from bias. Another significant concern revolves around the potential impact of AI language models on employment. As these technologies continue to advance, there's a possibility that they may replace human labor in various roles, such as customer support or data

entry. This could lead to job losses and economic instability, particularly among individuals with lower skill levels. It's imperative for policymakers and corporations to collaborate in ensuring that the benefits of AI technology are distributed fairly and equitably across society.

References

1. Montazeri Z, Niknam T, Aghaei J, Malik OP, Dehghani M, Dhiman G (2023) Golf optimization algorithm: a new game-based metaheuristic algorithm and its application to energy commitment problem considering resilience. Biomimetics 8(5):386

2. Liu Y, Ding H, Wang Z, Jin G, Li B, Yang Z, Dhiman G (2023) A chaos-based adaptive equilibrium optimizer algorithm for solving global optimization problems. Math Biosci Eng 20(9):17242–17271

3. Alferaidi A, Yadav K, Yasmeen S, Alharbi Y, Viriyasitavat W, Dhiman G, Kaur A (2023) Node multi-attribute network community healthcare detection based on graphical matrix factorization. J Circ Syst Comput 33:2450080

4. Kumar A, Misra R, Singh T, Dhiman G (2023) Apo-an feature selection based glo- rot init optimal transcnn landslide detection from multi source satellite imagery. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17090-2

5. Mekala M, Dhiman G, Park JH, Jung H-Y, Viriyasitavat W (2023) Asxc ^{2} approach: a service-x cost optimization strategy based on edge orchestration for iiot. IEEE Trans Industrial Inf. https://doi.org/10.1109/TII.2023.3315744

6. Garg RK, Soni SK, Vimal S, Dhiman G (2023) 3-d spatial correlation model for reducing the transmitting nodes in densely deployed wsn. Microprocess Microsyst 103:104963

7. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. OpenAI

8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser LU, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates Inc, Scotland

9. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert- Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates Inc, Scotland, pp 1877–1901

10. Rajpoot PK, Parikh A (2023) GPT-FinRE: In-context learning for financial relation extraction using large language models

11. Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020) How does NLP benefit legal system: a summary of legal artificial intelligence

12. Daniel Ruby: ChatGPT Statistics. https://www.demandsage.com/chatgpt-statistics (last accesses August 23, 2023)

13. Liu Y, Ding H, Wang Z, Dhiman G, Yang Z, Hu P (2023) An enhanced equilibrium optimizer for solving optimization tasks. Comput Mater Continua. https://doi.org/10.32604/cmc.2023.039883

14. Natarajan S, Sampath P, Arunachalam R, Shanmuganathan V, Dhiman G, Chakrabarti P, Chakrabarti T, Margala M (2023) Early diagnosis and meta-agnostic model visualization of tuberculosis based on radiography images. Sci Rep 13(1):22803

15. Chopra G, Rani S, Viriyasitavat W, Dhiman G, Kaur A, Vimal S (2024) Uavassisted partial co-operative noma based resource allocation in c2vx and tinyml based use case scenario. IEEE Internet J. https://doi.org/10.1109/JIOT.2024.3351733

16. Baba SM, Bala I, Dhiman G, Sharma A, Viriyasitavat W (2024) Automated diabetic retinopathy severity grading using novel dr-resnet+ deep learning model. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-18434-2

17. Dhiman G, Alghamdi NS (2024) Smose: Artificial intelligence-based smart city framework using multi-objective and iot approach for consumer electronics application. IEEE Trans Consumer Electron. https://doi.org/10.1109/TCE.2024.3363720

18. Sharma S, Gupta K, Gupta D, Rani S, Dhiman G (2024) An insight survey on sensor errors and fault detection techniques in smart spaces. CMES-Comput Model Eng Sci. https://doi.org/10.32604/cmes.2023.029997

19. Tegos AA, Tegos SA, Tyrovolas D, Diamantoulakis PD, Sarigiannidis P, Karagiannidis GK (2024) Breaking orthogonality in uplink with randomly deployed sources. IEEE Open J Commun Soc. https://doi.org/10.1109/OJCOMS.2023.3349181

20. Maghrabi LA, Shabanah S, Althaqafi T, Alsalman D, Algarni S, Abdullah A, Ragab M (2024) Enhancing cybersecurity in the internet of things environment using bald eagle search optimization with hybrid deep learning. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3352568

21. El Khoury R, Nasrallah N (eds) (2024) Intelligent systems, business, and innovation research. Springer International Publishing, Cham

22. Rabieinejad E, Yazdinejad A, Dehghantanha A, Srivastava G (2024) Two-level privacy-preserving framework: federated learning for attack detection in the consumer internet of things. IEEE Trans Consumer Electron. https://doi.org/10.1109/TCE.2024.3349490

23. Zhang D, Shafiq M, Srivastava G, Gadekallu TR, Wang L, Gu Z (2024) Stbciot: Securing the transmission of biometric images in customer iot. IEEE Internet J. https://doi.org/10.1109/JIOT.2024.3351988

24. Kozhaya SE, Haidar-Ahmad JA, Abdallah AA, Kassas ZM, Saab SS (2021) Comparison of neural network architectures for simultaneous tracking and navigation with leo satellites. In: Proceedings of the 34th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+2021), pp. 2507–2520

25. Saab S Jr, Fu Y, Ray A, Hauser M (2022) A dynamically stabilized recurrent neural network. Neural Process Lett 54(2):1195–1209

26. Sayour MH, Kozhaya SE, Saab SS (2022) Autonomous robotic manipulation: real-time, deep-learning approach for grasping of unknown objects. J Robotics. https://doi.org/10.1155/2022/2585656

27. Saab S Jr, Saab K, Phoha S, Zhu M, Ray A (2022) A multivariate adaptive gradient algorithm with reduced tuning efforts. Neural Netw 152:499–509

28. Saab S Jr, Phoha S, Zhu M, Ray A (2022) An adaptive polyak heavy-ball method. Mach Learn 111(9):3245–3277

29. Osone H, Lu J-L, Ochiai Y (2021) Buncho: Ai supported story co-creation via unsupervised multitask learning to increase writers' creativity in Japanese. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.

CHI EA '21. Association for Computing Machinery, New York. https://doi.org/10.1145/3411763.3450391

30. Baktash JA, Dawodi M (2023) Gpt-4: a review on advancements and opportunities in natural language processing

31. Ray PP (2023) Chatgpt: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Cyber-Phys Syst 3:121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

32. Zhang Z, Jiang H, Shen D, Saab SS (2023) Data-driven learning control algo- rithms for unachievable tracking problems. IEEE/CAA J Automatica Sinica. https://doi.org/10.1109/JAS.2023.123756

33. Kfouri R (2023) A robust deep learning approach for distribution system state esti- mation with distributed generation. PhD thesis, Lebanese American University (2023)

34. Tarhini A, Danach K, Harfouche A (2022) Swarm intelligence-based hyper-heuristic for the vehicle routing problem with prioritized customers. Ann Operations Res. https://doi.org/10.1007/s10479-020-03625-5

35. Abbas N, Nasser Y, Shehab M, Sharafeddine S (2021) Attack-specific feature selection for anomaly detection in software-defined networks. In: 2021 3rd IEEE Middle East and North Africa Communications Conference (menacomm), pp.142–146. IEEE

36. Gerges F, Shih F, Azar D (2021) Automated diagnosis of acne and rosacea using convolution neural networks. In: Proceedings of the 2021 4th International Conference on Artificial Intelligence and Pattern Recognition, pp. 607–613

37. Helwan A, Maaitah MKS, Uzelaltinbulat S, Altobel MZ, Darwish M (2021) Gaze prediction based on convolutional neural network. International Conference on Emerging Technologies and Intelligent Systems. Springer, Cham, pp 215–224

38. Yunis M, Markarian C, El-Kassar A (2020) A conceptual model for sustainable adoption of ehealth: Role of digital transformation culture and healthcare provider's readiness. Proceedings of the IMCIC 17

39. Haraty RA, Boukhari B, Kaddoura S (2021) An effective hash-based assessment and recovery algorithm for healthcare systems. Arabian J Sci Eng. https://doi.org/10.1007/s13369-021-06009-4

40. Arafeh M, El Barachi M, Mourad A, Belqasmi F (2019) A block-chain based archi- tecture for the detection of fake sensing in mobile crowdsensing. In: 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), pp.1–6. IEEE

41. Garg S, Peitz S, Nallasamy U, Paulik M (2019) Jointly learning to align and translate with transformer models

42. Jaques N, Ghandeharioun A, Shen JH, Ferguson C, Lapedriza A, Jones N, Gu S, Picard R (2019) Way off-policy batch deep reinforcement learning of implicit human preferences in dialog

43. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano PF, Leike J, Lowe R (2022) Training language models to follow instructions with human feedback. Adv Neural Inf Proc Syst 35:27730–27744

44. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (eds) (2022) Advances in neural information processing systems, vol 35. Curran Associates Inc, Scotland, pp 27730–27744

45. Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in NLP. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650. Association for Computational Linguistics, Florence, Italy. https://aclanthology.org/P19-1355

46. Liu H-I, Chen W-L (2022) X-transformer: A machine translation model enhanced by the self-attention mechanism. Appl Sci 12(9):4502. https://doi.org/10.3390/app12094502

47. Zhang Z, Strubell E, Hovy E (2023) A survey of active learning for natural language processing

48. Kenton JDM-WC, Toutanova LK (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, vol. 1, p. 2

49. Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, Christiano P, Irving G (2020) Fine-tuning language models from human preferences

50. Sorkhoh I, Ebrahimi D, Assi C, Sharafeddine S, Khabbaz M (2020) An infrastructure-assisted workload scheduling for computational resources exploitation in the fog-enabled vehicular network. IEEE Internet Things J 7(6):5021–5032

51. Khabbaz M, Assi C, Sharafeddine S (2021) Multihop v2u path availability analysis in UAV-assisted vehicular networks. IEEE Internet Things J 8(13):10745–10754

52. Rahman SA, Tout H, Talhi C, Mourad A (2020) Internet of things intrusion detection: centralized, on-device, or federated learning? IEEE Network 34(6):310–317

53. AbdulRahman S, Tout H, Mourad A, Talhi C (2020) Fedmccs: Multicriteria client selection model for optimal IOT federated learning. IEEE Internet Things J 8(6):4723–4735

54. Mourad A, Tout H, Wahab OA, Otrok H, Dbouk T (2020) Ad hoc vehicular fog enabling cooperative low-latency intrusion detection. IEEE Internet Things J 8(2):829–843

55. Chicha E, Bouna BA, Nassar M, Chbeir R, Haraty RA, Oussalah M, Benslimane D, Alraja MN (2021) A user-centric mechanism for sequentially releasing graph datasets under blowfish privacy. ACM Trans Internet Technol (TOIT) 21(1):1–25

56. Nour, C., Takche, J.: A general result about inner regularization of sets (2020)

57. Tarhini A, Harfouche A, De Marco M (2022) Artificial intelligence-based digital transformation for sustainable societies: the prevailing effect of covid-19 crises. Pacific Asia J Assoc Inf Sys 14(2):1

58. Tanwar S, Popat A, Bhattacharya P, Gupta R, Kumar N (2022) A taxonomy of energy optimization techniques for smart cities: architecture and future directions. Exp Syst 39(5):12703. https://doi.org/10.1111/exsy.12703

59. Henderson P, Hu J, Romoff J, Brunskill E, Jurafsky D, Pineau J (2020) Towards the systematic reporting of the energy and carbon footprints of machine learning. J Mach Learn Res 21(1):10039–10081

60. Schwartz R, Dodge J, Smith NA, Etzioni O (2019) Green ai. arxiv e-prints, art. arXiv preprint arXiv:1907.10597

61. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623

62. O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Nature Publishing Group, New York

63. Santos A, Aerle R, Barrientos L, Martinez-Urtaza J (2020) Computational methods for 16s metabarcoding studies using nanopore sequencing data. Comput Struct Biotechnol J 18:296–305

64. Verganti R, Vendraminelli L, Iansiti M (2020) Innovation and design in the age of artificial intelligence. J Prod Innov Manag 37(3):212–227

65. Latif E, Mai G, Nyaaba M, Wu X, Liu N, Lu G, Li S, Liu T, Zhai X (2023) Artificial general intelligence (AGI) for education

66. Luo L, Wang Y, Mo DY (2022) Identifying covid-19 personal health mentions from tweets using masked attention model. IEEE Access 10:59068–59077. https://doi.org/10.1109/ACCESS.2022.3179808

67. Mondal S, Das S, Vrana VG (2023) How to bell the cat? a theoretical review of generative artificial intelligence towards digital

disruption in all walks of life. Technologies. https://doi.org/10.3390/technologies11020044

68. Cooper G (2023) Examining science education in chatgpt: an exploratory study of generative artificial intelligence. J Sci Educ Technol 32(3):444–452

69. Haraty RA, Mansour N, Zeitunlian H (2018) Metaheuristic algorithm for state- based software testing. Appl Artif Intell 32(2):197–213

70. Hussain M, Kaassamani S, Auguste T, Boutu W, Gauthier D, Kholodtsova M, Gomes J-T, Lavoute L, Gaponov D, Ducros N et al (2021) Spectral control of high order harmonics through non-linear propagation effects. Appl Phys Lett. https://doi.org/10.1063/5.0053152

71. Bouri E, Gupta R, Wang S (2022) Nonlinear contagion between stock and real estate markets: International evidence from a local Gaussian correlation approach. Int J Financ Econ 27(2):2089–2109

72. Saab SS, Saab KK (2019) Shuffled linear regression with erroneous observations. In: 2019 53rd Annual Conference on Information Sciences and Systems (CISS), pp. 1–6. IEEE

73. Kassis MT, Tannir D, Toukhtarian R, Khazaka R (2019) Moments-based sensitivity analysis of x-parameters with respect to linear and nonlinear circuit components. In: 2019 IEEE 28th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS), pp. 1–3. IEEE

74. Chamoun S, Nour C (2021) A nonlinear φ0-convexity result for the bilateral minimal time function

75. Tiwari AK, Bathia D, Bouri E, Gupta R (2021) Investor sentiment connectedness: evidence from linear and nonlinear causality approaches. Ann Financial Econom 16(04):2150016

76. Shoja MM, Ridder JM, Rajput V (2023) The emerging role of generative artificial intelligence in medical education, research, and practice. Cureus 15(6):e40883

77. Bhattacharya P, Obaidat MS, Savaliya D, Sanghavi S, Tanwar S, Sadaun B (2022) Metaverse assisted telesurgery in healthcare 5.0: An interplay of blockchain and explainable ai. In: 2022 International Conference on Computer, Information and Telecommunication Systems (CITS), Piraeus, Greece, pp. 1–5. https://doi.org/10.1109/CITS55221.2022.9832978

78. Lv Z (2023) Generative artificial intelligence in the metaverse era. Cognitive Robotics

79. Gozalo-Brizuela R, Garrido-Merchan EC (2023) Chatgpt is not all you need. a state of the art review of large generative ai models. arXiv preprint arXiv:2301.04655

80. Issa JS (2022) A nonlinear absorber for the reflection of travelling waves in bars. Nonlinear Dyn 108(4):3279–3295

81. Marrouche W, Farah R, Harmanani HM (2018) A multiobjective optimization method for the soc test time, tam, and power optimization using a strength pareto evolutionary algorithm. In: Information Technology-New Generations: 14th International Conference on Information Technology, pp. 685–695. Springer, Cham

82. Abbas N, Fawaz W, Sharafeddine S, Mourad A, Abou-Rjeily C (2022) Svm-based task admission control and computation offloading using lyapunov optimization in heterogeneous MEC network. IEEE Trans Netw Serv Manage 19(3):3121–3135

83. Ghanem CR, Gereige EN, Bou Nader WS, Mansour CJ (2022) Stirling system optimization for series hybrid electric vehicles. J Automob Eng 236(2–3):407–423

84. Abboud R, Tekli J (2020) Integration of nonparametric fuzzy classification with an evolutionary-developmental framework to perform music sentiment-based analysis and composition. Soft Comput 24(13):9875–9925

85. Salloum G, Tekli J (2021) Automated and personalized nutrition health assessment, recommendation, and progress evaluation using fuzzy reasoning. Int J Hum Comput Stud 151:102610

86. Kouatli I (2018) Fuzzimetric employee evaluations system (fees): a multivariable- modular approach. J Intell Fuzzy Syst 35(4):4717–4729

87. Baidoo-Anu D, Owusu Ansah L (2023) Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. Available at SSRN 4337484

88. Xia L, Sun Y, Liang C, Zhang L, Imran MA, Niyato D (2023) Generative ai for semantic communication: Architecture, challenges, and outlook. arXiv preprint arXiv:2308.15483\

89. Taheri R, Kabuli M, Vryzas Z (2020) Fracturing and permeability enhancement with laser technology employing fuzzy logic. J Petrol Sci Eng 188:106830

90. Azadeh A, Kalantari M, Ahmadi G, Eslami H (2019) A flexible genetic algorithm- fuzzy regression approach for forecasting: the case of bitumen consumption. Constr Innov 19(1):71–88

91. Abdallah SB, Kouatli I (2018) Fuzzy volatility effect on major projects timing. In: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6. IEEE

92. Ben Abdallah S, Kouatli I (2020) Fuzzy volatility of project option value based on trapezoidal membership functions. In: Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making: Proceedings of the INFUS 2019 Conference, Istanbul, Turkey, July 23–25, 2019, pp. 1307–1314. Springer, Cham

93. Kouatli I (2020) The use of fuzzy logic as augmentation to quantitative analysis to unleash knowledge of participants' uncertainty when filling a survey: case of cloud computing. IEEE Trans Knowl Data Eng 34(3):1489–1500

94. Saab SS, Shen D, Orabi M, Kors D, Jaafar RH (2021) Iterative learning con- trol: Practical implementation and automation. IEEE Trans Ind Electron 69(2):1858–1866

95. Saab SS, Jaafar RH (2021) A proportional-derivative-double derivative controller for robot manipulators. Int J Control 94(5):1273–1285

96. Shen D, Huo N, Saab SS (2021) A probabilistically quantized learning control framework for networked linear systems. IEEE Trans Neural Networks Learn Syst 33(12):7559–7573

97. Shen D, Saab SS (2021) Noisy-output-based direct learning tracking control with Markov nonuniform trial lengths using adaptive gains. IEEE Trans Autom Control 67(8):4123–4130

98. Prasad VK, Bhavsar MD, Tanwar S (2019) Influence of montoring: fog and edge computing. Scalable Comput 20(2):365–376

99. George AS, George AH, Baskar T, Martin AG (2023) Revolutionizing business communication: exploring the potential of gpt-4 in corporate settings. Partners Univ Int Res J 2(1):149–157

100. Zhang C, Zhang C, Zheng S, Qiao Y, Li C, Zhang M, Dam SK, Thwal CM, Tun YL, Huy LL et al. (2023) A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need? arXiv preprint arXiv:2303.11717

101. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X (2020) Pre-trained models for natural language processing: a survey. Sci China Technol Sci 63(10):1872–1897

102. Lund BD, Wang T (2023) Chatting about chatgpt: how may ai and gpt impact academia and libraries? Lib Hi Tech News 40(3):26–29

103. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E et al (2023) Chatgpt for good? On opportunities and challenges of large language models for education. Learn Ind Diff 103:102274

104. Chen L, Zaharia M, Zou J (2023) FrugalGPT: how to use large language models while reducing cost and improving performance

105. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR

106. Li L, Wang X, Wang K, Lin Y, Xin J, Chen L, Xu L, Tian B, Ai Y, Wang J et al (2019) Parallel testing of vehicle intelligence via virtual-real interaction. Sci Robot 4(28):4106

107. Li T, Sahu AK, Talwalkar A, Smith V (2020) Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 37(3):50–60

108. He D, Xia Y, Qin T, Wang L, Yu N, Liu T-Y, Ma W-Y (2016) Dual learning for machine translation. Adv Neural Information Proc Syst. https://doi.org/10.48550/arXiv.1611.00179

109. Wang X, Chen Y, Zhu W (2021) A survey on curriculum learning. IEEE Trans Pattern Anal Mach Intell 44(9):4555–4576

110. Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160–167

111. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76

112. Tian Y, Zhao X, Huang W (2022) Meta-learning approaches for learning-to-learn in deep learning: a survey. Neurocomputing 494:203–223

113. Chen Z, Liu B (2018) Lifelong machine learning. Synth Lect Artif Intell Mach Learn 12(3):1–207

114. Mienye ID, Sun Y (2022) A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access 10:99129–99149

115. Becker S, Hinton GE (1992) Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355(6356):161–163

116. Brendan McMahan H, Moore E, Ramage D, Hampson S, Arcas B (206) Communication-efficient learning of deep networks from decentralized data. arXiv e-prints, 1602

117. Bhattacharya P, Tiwari AK, Singh A (2023) Dual-buffer-based optical datacenter switch design. J Optic Commun 44(2):155–162. https://doi.org/10.1515/joc-2019-0023

118. Verma A, Bhattacharya P, Bodkhe U, Ladha A, Tanwar S (2021) Dams: dynamic association for view materialization based on rule mining scheme. In: Singh PK, Singh Y, Kolekar MH, Kar AK, Chhabra JK, Sen A (eds) Recent innovations in computing. Springer, Singapore, pp 529–544

119. Prasad VK, Bhavsar MD (2020) Monitoring and prediction of SLA for IOT based cloud. Scalable Comput 21(3):349–358

120. Parmar J, Sanghavi S, Prasad V, Shah P (2023) Microservice architecture observ- ability tool analysis. In: Reddy VS, Prasad VK, Wang J, Reddy KTV (eds) Soft computing and signal processing. Springer, Singapore, pp 1–8

121. Prasad VK, Bhavsar MD (2020) Monitoring iaas cloud for healthcare systems: Healthcare information management and cloud resources utilization. Int J e-Health Med Commun (IJE-HMC) 11(3):54–70

122. Verma A, Bhattacharya P, Bodkhe U, Saraswat D, Tanwar S, Dev K (2023) Fedrec: Trusted rank-based recommender scheme for service provisioning in federated cloud environment. Digital Commun Networks 9(1):33–46. https://doi.org/10.1016/j.dcan.2022.06.003

123. Bhattacharya P, Patel SB, Gupta R, Tanwar S, Rodrigues JJPC (2022) Satya: Trusted bi-lstm-based fake news classification scheme for smart community. IEEE Trans Comput Soc Syst. https://doi.org/10.1109/TCSS.2021.3131945

124. Prasad VK, Tanwar S, Bhavsar MD (2021) Advance cloud data analytics for 5G enabled IoT. Springer, Cham, pp 159–180

125. Verma A, Bhattacharya P, Budhiraja I, Gupta AK, Tanwar S (2022) Fusion of federated learning and 6g in internet-of-medical-things Architecture case study and emerging directions. In: Chhabra JK, Tanwar S, Singh PK, Wierzchon ST (eds)

126. Bhattacharya P, Bodkhe U, Zuhair M, Rashid M, Liu X, Verma A, Kishan Dewangan R (2021) Amalgamation of blockchain and sixth-generation- envisioned responsive edge orchestration in future cellular vehicle-to-anything ecosystems: opportunities and challenges. Trans Emerging Telecommun Technol. https://doi.org/10.1002/ett.4410

127. Bhattacharya P, Singh A, Kumar A, Tiwari AK, Srivastava R (2017) Com- parative study for proposed algorithm for all-optical network with negative acknowledgement (ao-nack). In: Proceedings of the 7th International Conference on Computer and Communication Technology. ICCCT-2017, pp. 47–51. Association for Computing Machinery, New York. https://doi.org/10.1145/3154979.3154981 .

128. Saraswat D, Verma A, Bhattacharya P, Tanwar S, Sharma G, Bokoro PN, Sharma R (2022) Blockchain-based federated learning in UAVs beyond 5g net- works: a solution taxonomy and future directions. IEEE Access 10:33154–33182. https://doi.org/10.1109/ACCESS.2022.3161132

129. Verma A, Bhattacharya P, Zuhair M, Tanwar S, Kumar N (2022) Vacochain: Blockchain-based 5g-assisted uav vaccine distribution scheme for future pan- demics. IEEE J Biomed Health Inform 26(5):1997–2007. https://doi.org/10.1109/JBHI.2021.3103404

130. Verma A, Bhattacharya P, Saraswat D, Tanwar S, Kumar N, Sharma R (2023) Sanjeevni: Secure UAV-envisioned massive vaccine distribution for covid-19 underlying 6g network. IEEE Sens J 23(2):955–968. https://doi.org/10.1109/JSEN.2022.3188929

131. Saraswat D, Bhattacharya P, Singh A, Verma A, Tanwar S, Kumar N (2022) Secure 5g-assisted UAV access scheme in IOBT for region demarcation and surveillance operations. IEEE Commun Stand Mag 6(1):58–66. https://doi.org/10.1109/MCOMSTD.0001.2100057

132. Prasad VK, Bhavsar M (2018) Efficient resource monitoring and prediction techniques in an IAAS level of cloud computing: Survey. In: Patel Z, Gupta S (eds) Future internet technologies and trends. Springer, Cham, pp 47–55

133. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z (2023) Chatgpt and a new academic reality: artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. J Am Soc Inf Sci 74(5):570–581

134. Lo LS (2023) The clear path: a framework for enhancing information literacy through prompt engineering. J Acad Librariansh 49(4):102720. https://doi.org/10.1016/j.acalib.2023.102720

135. Hu X, Tian Y, Nagato K, Nakao M, Liu A (2023) Opportunities and challenges of ChatGPT for design knowledge management

136. Hosseini M, Horbach SP (2023) Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. Res Integrity Peer Rev 8(1):1–9

137. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S (2023) Chatgpt vs google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res 25:48966. https://doi.org/10.2196/48966

138. Chu MN (2023) Assessing the benefits of chatgpt for business: an empirical study on organizational performance. IEEE Access 11:76427–76436. https://doi.org/10.1109/ACCESS.2023.3297447

139. Ausat AMA, Azzaakiyyah HK, Permana RM, Riady Y, Suherlan S (2023) The role of chatgpt in enabling msmes to compete in the digital age. Innovative 3(2):622–631. https://doi.org/10.31004/innovative.v3i2.346

140. Rijcken E, Scheepers F, Zervanou K, Spruit M, Mosteiro P, Kaymak U (2023) Towards Interpreting Topic Models with

ChatGPT. In: The 20th World Congress of the International Fuzzy Systems Association, IFSA (2023). The 20th World Congress of the International Fuzzy Systems Association, IFSA ; Conference date: 20-08-2023 Through 24-08-2023. https://ifsa2023.org/

141. Cox A, Seth I, Xie Y, Hunter-Smith DJ, Rozen WM (2023) Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients. Aesthetic Surg J 43(8):658–662. https://doi.org/10.1093/asj/sjad096

142. Wang F-Y, Li J, Qin R, Zhu J, Mo H, Hu B (2023) Chatgpt for computational social systems: from conversational applications to human-oriented operating systems. IEEE Trans Comput Soc Syst 10(2):414–425. https://doi.org/10.1109/TCSS.2023.3252679

143. Panda S, Kaur N (2023) Exploring the viability of chatgpt as an alternative to traditional chatbot systems in library and information centers. Lib Hi Tech News 40(3):22–25

144. Lo CK (2023) What is the impact of chatgpt on education? a rapid review of the literature. Educ Sci 13(4):410. https://doi.org/10.3390/educsci13040410

145. Abdullayeva M, Musayeva ZM (2023) The impact of chat gpt on student's writing skills: An exploration of ai-assisted writing tools. In: International Conference of Education, Research and Innovation, vol. 1, pp. 61–66

146. Ausat AMA, Rachman A, Rijal S, Suherlan S, Azzaakiyyah HK (2023) Application of chatgpt in improving operational efficiency in the context of entrepreneurship. Jurnal Minfo Polgan 12(1):1220–1228

147. Koonchanok R, Pan Y, Jang H (2023) Tracking public attitudes toward ChatGPT on Twitter using sentiment analysis and topic modeling

148. Javaid M, Haleem A, Singh RP (2023) Chatgpt for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Eval 3(1):100105. https://doi.org/10.1016/j.tbench.2023.100105

149. Verma A, Bhattacharya P, Madhani N, Trivedi C, Bhushan B, Tanwar S, Sharma G, Bokoro PN, Sharma R (2022) Blockchain for industry 5.0: vision, opportunities, key enablers, and future directions. IEEE Access 10:69160–69199. https://doi.org/10.1109/ACCESS.2022.3186892

150. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, Bokoro PN, Sharma R (2022) Explainable ai for healthcare 5.0: opportunities and challenges. IEEE Access 10:84486–84517. https://doi.org/10.1109/ACCESS.2022.3197671

151. Welch S (1975) Comparative studies on the human glutamate-pyruvate transaminase phenotypes—gpt 1, gpt 2–1, gpt 2. Humangenetik 30:237–249

152. Dehouche N (2021) Plagiarism in the age of massive generative pre-trained trans- formers (gpt-3). Ethics Sci Environ Politics 21:17–23

153. Yue T, Au D, Au CC, Iu KY (2023) Democratizing financial knowledge with chatgpt by openai: Unleashing the power of technology. Available at SSRN 4346152

154. Ali JKM, Shamsan MAA, Hezam TA, Mohammed AAQ (2023) Impact of chatgpt on learning motivation: teachers and students' voices. J English Stud Arabia Felix 2(1):41–49. https://doi.org/10.56540/jesaf.v2i1.51

155. Surameery NMS, Shakor MY (2023) Use chat gpt to solve programming bugs. Int J Inf Technol. https://doi.org/10.55529/ijitc.31.17.22

156. Pavlik JV (2023) Collaborating with chatgpt: considering the implications of gen- erative artificial intelligence for journalism and media education. J Mass Commun Edu 78(1):84–93. https://doi.org/10.1177/10776958221149577

157. Ali H, Aysan AF (2023) What will chatgpt revolutionize in financial industry? Available at SSRN 4403372

158. Iftikhar L, Iftikhar M, Hanif M (2023) Docgpt: Impact of chatgpt-3 on health services as a virtual doctor. EC Paediatrics 12:45–55

159. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepanˆo C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J et al (2023) Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. PLoS Digital Health 2(2):0000198

160. Haleem A, Javaid M, Singh RP (2022) An era of chatgpt as a significant futur- istic support tool: a study on features, abilities, and challenges. BenchCouncil Trans Benchmarks Stand Eval 2(4):100089. https://doi.org/10.1016/j.tbench.2023.100089

161. Alexakis G, Panagiotakis S, Fragkakis A, Markakis E, Vassilakis K (2019) Control of smart home operations using natural language processing, voice recognition and IOT technologies in a multi-tier architecture. Designs 3(3):32

162. Aydın O, Karaarslan E (2022) Openai chatgpt generated literature review: Digital twin in healthcare. Available at SSRN 4308687

163. Wang F-Y, Yang J, Wang X, Li J, Han Q-L (2023) Chat with chatgpt on industry 5.0: learning and decision-making for intelligent industries. IEEE/CAA J Automatica Sinica 10(4):831–834. https://doi.org/10.1109/JAS.2023.123552

164. Deng J, Lin Y (2022) The benefits and challenges of chatgpt: an overview. Front Comput Intell Syst 2(2):81–83

165. Mijwil M, M Hiran KK, Doshi R, Dadhich M, Al-Mistarehi A-H, Bala I (2023) Chatgpt and the future of academic integrity in the artificial intelligence era: a new frontier. Al-Salam J Eng Technol 2(2):116–127. https://doi.org/10.55145/ajest.2023.02.02.015

166. Gill SS, Kaur R (2023) Chatgpt: vision and challenges. Internet Cyber-Phys Syst 3:262–271

167. Bhattacharya P, Patel K, Zuhair M, Trivedi C (2022) A lightweight authenti- cation via unclonable functions for industrial internet-of-things. In: 2022 2nd International Conference on Innovative Practices in Technology and Manage- ment (ICIPTM), Gautam Buddha Nagar, India, vol. 2, pp. 657–662. https://doi.org/10.1109/ICIPTM54933.2022.9754198

168. Kirk HR, Jun Y, Volpin F, Iqbal H, Benussi E, Dreyer F, Shtedrit- ski, A., Asano, Y. (2021) Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW (eds) Advances in neural information processing systems, vol 34. Curran Asso- ciates Inc, Scotland, pp 2611–2624

169. Nozza D, Bianchi F, Hovy D (2021) HONEST: measuring hurtful sentence com- pletion in language models. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2398–2406. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.naacl-main.191 .

170. Schick T, Udupa S, Schütze H (2021) Self-diagnosis and self-debiasing: a proposal for reducing corpus-based bias in NLP. Trans Assoc Comput Linguis 9:1408–1424. https://doi.org/10.1162/tacla00434

171. Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang K-W, Gupta R (2021) Bold: dataset and metrics for measuring biases in open-ended language generation. In: Proceedings of the 2021 ACM Conference on Fair- ness, Accountability, and Transparency. FAccT '21, pp. 862–872. Association for Computing Machinery, New Yorks. https://doi.org/10.1145/3442188.3445924

172. Yeo C, Chen A (2020) Defining and evaluating fair natural language generation

173. Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, Singer Y, Shieber S (2020) Investigating gender bias in language models using

causal mediation analysis. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in neural information processing systems, vol 33. Curran Associates Inc., Scotland, pp 12388–12401

174. Sheng E, Chang K-W, Natarajan P, Peng N (2019) The Woman Worked as a Babysitter: On Biases in Language Generation

175. Solaiman I, Brundage M, Clark J, Askell A, Herbert-Voss A, Wu J, Radford A, Krueger G, Kim JW, Kreps S, McCain M, Newhouse A, Blazakis J, McGuffie K, Wang J (2019) Release strategies and the social impacts of language models

176. Qian Y, Muaz U, Zhang B, Hyun JW (2019) Reducing gender bias in word-level language models with a gender-equalizing loss function

177. Bordia S, Bowman SR (2019) Identifying and reducing gender bias in word-level language models

178. Sheng E, Chang K-W, Natarajan P, Peng N (2021) "Nice Try, Kiddo": Investi- gating Ad Hominems in Dialogue Responses

179. Cercas Curry A, Robertson J, Rieser V (2020) Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In: Proceed- ings of the Second Workshop on Gender Bias in Natural Language Processing, pp. 72–78. Association for Computational Linguistics, Barcelona, Spain (Online). https://aclanthology.org/2020.gebnlp-1.7

180. Dinan E, Fan A, Wu L, Weston J, Kiela D, Williams A (2020) Multi- dimensional gender bias classification

181. Henderson P, Sinha K, Angelard-Gontier N, Ke NR, Fried G, Lowe R, Pineau J (2018) Ethical challenges in data-driven dialogue systems. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18, pp. 123–129. Association for Computing Machinery, New York. https://doi.org/10.1145/3278721.3278777

182. Tomalin M, Byrne B, Concannon S, Saunders D, Ullmann S (2021) The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. Ethics Inf Technol. https://doi.org/10.1007/s10676-021-09583-1

183. Saunders D, Sallis R, Byrne B (2022) First the worst: Finding better gender translations during beam search

184. Choubey PK, Currey A, Mathur P, Dinu G (2021) Improving gender translation accuracy with filtered self-training

185. Renduchintala A, Williams A (2021) Investigating failures of automatic translation in the case of unambiguous gender

186. Savoldi B, Gaido M, Bentivogli L, Negri M, Turchi M (2021) Gender bias in machine translation. Trans Assoc Comput Linguist 9:845–874. https://doi.org/10.1162/tacla00401

187. Cho WI, Kim J, Yang J, Kim NS (2021) Towards cross-lingual generalization of translation gender bias. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 449–457. Association for Computing Machinery, New York. https://doi.org/10.1145/3442188.3445907

188. Roberts N, Liang D, Neubig G, Lipton ZC (2020) Decoding and diversity in machine translation

189. Hovy D, Bianchi F, Fornaciari T (2020) "You sound just like your father" com- mercial machine translation systems include stylistic biases. In: Proceedings of the 58th Annual Meeting of the Association for Computational Lin- guistics, pp. 1686–1690. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.154

190. Stafanovičs A, Bergmanis T, Pinnis M (2020) Mitigating gender bias in machine translation with target gender annotations

191. Ferrer X, Nuenen T, Such JM, Criado N (2021) Discovering and categorising language biases in reddit. Proc Int AAAI Conf Web Soc Med 15(1):140–151. https://doi.org/10.1609/icwsm.v15i1.18048

192. Basta CRS, Ruiz Costa-Jussà M, Rodríguez Fonollosa JA (2020) Towards mit- igating gender bias in a decoder-based neural machine translation model by adding contextual information. In: Proceedings of the The Fourth Widening Natural Language

193. Costa-jussa MR, Escolano C, Basta C, Ferrando J, Batlle RKharitonova K (2020) Gender bias in multilingual neural machine translation: the architecture matters

194. Costa-jussa MR, Jorge A (2020) Fine-tuning neural machine translation on gender- balanced datasets. In: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pp. 26–34. Association for Computational Linguistics, Barcelona, Spain (Online). https://aclanthology.org/2020. gebnlp-1.3

195. Kocmi T, Limisiewicz T, Stanovsky G (2020) Gender Coreference and Bias Evaluation at WMT 2020

196. Saunders D, Byrne B (2020) Reducing gender bias in neural machine translation as a domain adaptation problem

197. Moryossef A, Aharoni R, Goldberg Y (2019) Filling gender & number gaps in neural machine translation with black-box context injection

198. Font JE, Costa-jussa MR (2019) Equalizing gender biases in neural machine translation with word embeddings techniques

199. Rescigno AA, Vanmassenhove E, Monti J, Way A (2020) A case study of natural gender phenomena in translation a comparison of google translate, Bing Microsoft translator and Deepl for English to Italian, French and Spanish. Comput Linguistics CLiC-it 2020:359

200. Elaraby M, Tawfik AY, Khaled M, Hassan H, Osama A (2018) Gender aware spoken language translation applied to english-arabic. In: 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), pp. 1–6. https://doi.org/10.1109/ICNLSP.2018.8374387

201. Vanmassenhove E, Hardmeier C, Way A (2018) Getting gender right in neural machine translation. In: Proceedings of the 2018 Conference on Empirical Meth- ods in Natural Language Processing. Association for Computational Linguistics. https://doi.org/10.18653/v1/d18-13342

202. Sun T, Webster K, Shah A, Wang WY, Johnson M (2021) They, them, theirs: rewriting with gender-neutral English (2021)

203. Alhafni B, Habash N, Bouamor H (2020) Gender-aware reinflection using linguis- tically enhanced neural models. In: Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pp. 139–150. Association for Com- putational Linguistics, Barcelona

204. Zmigrod R, Mielke SJ, Wallach H, Cotterell R (2020) Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology

205. Habash N, Bouamor H, Chung C (2019) Automatic gender identification and rein- flection in Arabic. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp. 155–165. Association for Computational Lin- guistics, Florence. https://doi.org/10.18653/v1/W19-3822

206. Huang P-S, Zhang H, Jiang R, Stanforth R, Welbl J, Rae J, Maini V, Yogatama D, Kohli P (2020) Reducing sentiment bias in lan- guage models via counterfactual evaluation

207. Groenwold S, Ou L, Parekh A, Honnavalli S, Levy S, Mirza D, Wang WY (2020) Investigating African-American vernacular English in transformer-based text generation

208. Abid A, Farooqi M, Zou J (2021) Persistent Anti-Muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES '21, pp. 298–306. Association for Computing Machinery, New York. https://doi.org/10.1145/3461702.3462624

209. Sheng E Chang K-W, Natarajan P, Peng N (2020) Towards con- trollable biases in language generation

210. Peng X, Li S, Frazier S, Riedl M (2020) Reducing non-normative text generation from language models

211. Shwartz V, Rudinger R, Tafjord O. (2020) "You are grounded!": Latent name artifacts in pre-trained language models

212. Sheng E, Arnold J, Yu Z, Chang K-W, Peng N (2021) Revealing persona biases in dialogue systems

213. Ma X, Sap M, Rashkin H, Choi Y (2020) Powertransformer: unsupervised controllable revision for biased language correction

214. Pryzant R, Diehl Martinez R, Dass N, Kurohashi S, Jurafsky D, Yang D (2020) Automatically neutralizing subjective bias in text. Proc AAAI Conf Artif Intell 34(01):480–489. https://doi.org/10.1609/aaai.v34i01.5385

215. Shyamsukha S, Bhattacharya P, Patel F, Tanwar S, Gupta R, Pricop E (2021) Porf: Proof-of-reputation-based consensus scheme for fair transaction ordering. In: 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania, pp. 1–6. https://doi.org/10.1109/ECAI52376.2021.9515090

216. Khowaja SA, Khuwaja P, Dev K (2023) ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review

217. Stahl BC, Eke D (2024) The ethics of chatgpt – exploring the ethical issues of an emerging technology. Int J Inf Manage 74:102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

218. Huallpa JJ et al (2023) Exploring the ethical considerations of using chat GPT in university education. Periodicals Eng Nat Sci 11(4):105–115

219. Balamurugan R, Mohite S, Raja S (2023) Protein sequence classification using bidirectional encoder representations from transformers (Bert) approach. SN Comput Sci 4(5):481

220. Rasul T, Nair S, Kalendra D, Robin M, Oliveirasantini F, Ladeira WJ, Sun M, Day I, Rather RA, Heathcote L (2023) The role of chatgpt in higher education: benefits, challenges, and future research directions. J Appl Learn Teach. https://doi.org/10.37074/jalt.2023.6.1.29

221. Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, Chang JP, Kuo JS, Su KP (2023) The now and future of chatgpt and GPT in psychiatry. Psychiatry Clin Neurosci 77(11):592–6

222. Open AI Platform: LLM API. https://platform.openai.com/. Accessed: 2023-10-15

223. AI21 Platform: LLM API. https://www.ai21.com/. Accessed: 2023-10-15

224. TextSynth Platform: LLM API. https://textsynth.com/. Accessed: 2023-10-15

225. TextSynth Platform: LLM API. https://cohere.com/. Accessed: 2023–10–15

226. Zheng L, Guha N, Anderson BR, Henderson P, Ho DE (2021) When Does Pre-training Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset

227. Reddy S, Chen D, Manning CD (2019) CoQA: A conversational question answering challenge

228. Kumar A, Kumar M, Mahapatra RP, Bhattacharya P (2023) Le T-T-H, Verma S, Kavita Mohiuddin K (2023) Flamingo-optimization-based deep convolutional neural network for IOT-based arrhythmia classification. Sensors. https://doi.org/10.3390/s23094353

Authors and Affiliations

Pronaya Bhattacharya[1] · Vivek Kumar Prasad[2] · Ashwin Verma[2] · Deepak Gupta[3] · Assadaporn Sapsomboon[4] · Wattana Viriyasitavat[4] · Gaurav Dhiman[5,6,7,8,9,10]

✉ Gaurav Dhiman
hardgolf@gmail.com

Pronaya Bhattacharya
pbhattacharya@kol.amity.edu

Vivek Kumar Prasad
vivekprasad@nirmauni.ac.in

Ashwin Verma
ashwin.verma@nirmauni.ac.in

Deepak Gupta
drdeepakgupta.cse@gmail.com

Assadaporn Sapsomboon
assadaporn@cbs.chula.ac.th

Wattana Viriyasitavat
assadaporn@cbs.chula.ac.th

[1] Department of Computer Science and Engineering, Amity Schoolof Engineering and Technology, and Research and Innovation Cell, Amity University, Kolkata, West Bengal 700135, India

[2] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat 382481, India

[3] Department of Computer Science and Engineering, Maharaja Agarasen Institute of Technology, Delhi, India

[4] Business Information Technology Division, Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, Thailand

[5] Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon

[6] Department of Computer Science and Engineering, Graphic Era Deemed to Bee University, Dehradun, India

[7] Division of Research and Development, Lovely Professional University, Phagwara, India

[8] Centre of Research Impact and Outreach, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

[9] MEU Research Unit, Middle East University, Amman, Jordan

[10] School of Sciences and Emerging Technologies, Jagat Guru Nanak Dev Punjab State Open University, Patiala 147001, India