A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage

Muhammad Usman Hadi[1,*], Qasem Al-Tashi[2,*], Rizwan Qureshi[2,*], Abbas Shah[3], Amgad Muneer[2], Muhammad Irfan[4], Anas Zafar[5], Muhammad Bilal Shaikh[6], Naveed Akhtar[7], Jia Wu[2], and Seyedali Mirjalili[8,9]

[1]*School of Engineering, Ulster University, Belfast, BT15 1AP, United Kingdom (m.hadi@ulster.ac.uk)*
[2]*Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA (qaal@mdanderson.org; frizwan@mdanderson.org; amabdulraheem@mdanderson.org; JWu11@mdanderson.org)*
[3]*Department of Electronics Engineering, Mehran University of Engineering and Technology, Jamshoro,76062 Pakistan (abbasshah.syed@gmail.com)*
[4]*Faculty of Electrical Engineering, Ghulam Ishaq Khan Institute (GIKI) of Engineering Sciences and Technology, Swabi, 23460 Pakistan (mirfan@giki.edu.pk)*
[5]*Department of Computer Science, National University of Computer and Emerging Sciences, Karachi, Pakistan (anaszafar98@gmail.com)*
[6]*Center for Artificial Intelligence and Machine Learning (CAIML), Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Perth, Australia (mbshaikh@our.ecu.edu.au)*
[7]*University of Western Australia, 35 Stirling Hwy, Crawley WA 6009, Australia*
[8]*Centre for Artificial Intelligence Research and Optimization, Torrens University Australia, Fortitude Valley, Brisbane, QLD 4006, Australia (ali.mirjalili@torrens.edu.au)*
[9]*University Research and Innovation Center, Obuda University, 1034 Budapest, Hungary*

Abstract

Within the vast expanse of computerized language processing, a revolutionary entity known as Large Language Models (LLMs) has emerged, wielding immense power in its capacity to comprehend intricate linguistic patterns and conjure coherent and contextually fitting responses. Large language models (LLMs) are a type of artificial intelligence (AI) that have emerged as powerful tools for a wide range of tasks, including natural language processing (NLP), machine translation, and question-answering. This survey paper provides a comprehensive overview of LLMs, including their history, architecture, training methods, applications, and challenges. The paper begins by discussing the fundamental concepts of generative AI and the architecture of generative pre-trained transformers (GPT). It then provides an overview of the history of LLMs, their evolution over time, and the different training methods that have been used to train them. The paper then discusses the wide range of applications of LLMs, including medical, education, finance, and engineering. It also discusses how LLMs are shaping the future of AI and how they can be used to solve real-world problems. The paper then discusses the challenges associated with deploying LLMs in real-world scenarios, including ethical considerations, model biases, interpretability, and computational resource requirements. It also highlights techniques for enhancing the robustness and controllability of LLMs and addressing bias, fairness, and generation quality issues. Finally, the paper concludes by highlighting the future of LLM research and the challenges that need to be addressed in order to make LLMs more reliable and useful. This survey paper is intended to provide researchers, practitioners, and enthusiasts with a comprehensive understanding of LLMs, their evolution, applications, and challenges. By consolidating the state-of-the-art knowledge in the field, this survey serves as a valuable resource for further advancements in the development and utilization of LLMs for a wide range of real-world applications. The GitHub repo for this project is available at https://github.com/anas-zafar/LLM-Survey

Index Terms

Large Language Models, Generative AI, Conversational AI, Co-pilots, Natural language processing, GPT, ChatGPT, Bing, Bard, AI-enabled Tools, AI chatbots.

A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage

I. Introduction

Human beings possess the remarkable ability to express and communicate through language, which starts developing during early childhood and continues to evolve throughout their lifetime [1], [2]. However, machines lack the inherent capacity to comprehend and communicate in human language unless they are equipped with powerful AI algorithms [3]. The goal of achieving human-like reading, writing, and communication skills in machines has been a long-standing research challenge and desire [4] [5].

The advent of large language models (LLMs) can be attributed to advancements in deep learning (DL) methods, the availability of massive computational resources, and the availability of vast amounts of training data. These models, often pre-trained on extensive corpora from the web, possess the ability to learn complex patterns, linguistic nuances, and semantic relationships. Fine-tuning these models on specific downstream tasks has shown promising results, achieving state-of-the-art performance in various benchmarks [6].

Language modeling (LM) is a crucial approach to enhancing the language intelligence of machines. In general, LM involves modeling the probability of word sequences to predict the likelihood of the future. As can be seen from Fig 1, LM research has received widespread attention and has undergone four significant development stages, as follows: The first major development in LM was the Statistical Language Models (SLMs) [7], such as n-gram models [8]. These models estimate the likelihood of the next word in a sequence based on the frequency of occurrence of previous n-grams of words [9], [10]. For instance, a bigram model would use the frequency of occurrence of pairs of words to estimate the probability of the next word. The second stage of LM development involved the introduction of neural network-based LM, named Neural Language Models (NLMs) [11], which is also known as neural language modeling. This approach utilizes neural networks to predict the probability distribution of the next word in a sequence given the previous words in the sequence. Recurrent neural networks (RNNs) and their variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) are often used in this approach [12], [13]. The third stage of LM development involves the creation of contextualized word embeddings, which aim to capture the meaning and context of words in a sentence or text, called Pre-trained Language Models (PLMs). These models utilize neural networks to learn a vector representation of words that takes into account the context in which the word appears [14]. Examples of contextualized word embeddings include ELMo [15] and BERT [16]. In Table II, we present sources of data for pre-training LLaMA.

TABLE I: List of Acronyms and corresponding definitions.

Acronym	Definition
AI	Artificial Intelligence
AGI	Artificial General Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CV	Computer Vision
CTRL	Conditional Transformer Language Model
FFF	Fused Filament Fabrication
GANs	Generative Adversarial Networks
GNMT	Google Neural Machine Translation
GPT	Generative Pre-Trained transformers
GPT-3	Generative Pre-trained Transformer 3
GPT-4	Generative Pre-trained Transformer 4
GPUs	Graphical Processing Units
GRUs	Gated Recurrent Units
LLaMA	Large Language Model Meta AI
LLM	Large Language Models
LM	Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning
MLM	Masked Language Modeling
NSP	next-sentence prediction
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PLMs	Pre-trained Language Models
RNN	Recurrent neural networks
RNNLM	Recurrent neural network language model
SLMs	Statistical Language Models
T5	Text-to-Text Transfer Transformer
TPUs	Tensor Processing Units
USMLE	United States Medical Licensing Exam
VL-PTMs	Vision-Language Pre-trained Models
XLNet	eXtreme Language Understanding Network

The fourth stage of LM development involved the creation of large-scale pretraining language models called Large Language Models (LLMs) [17], which have the capability of performing various natural language processing (NLP) tasks with excellent performance [18]. These models, such as GPT-3 and GPT-4, are trained on massive amounts of text data and can be fine-tuned for specific downstream tasks, such as language translation or question answering [19].

In summary, these four stages of LM development represent major advancements in the field, with each stage building upon the previous one and pushing the boundaries of what machines can achieve in NLP and Computer Vision. In this research, we mainly focus on LLMs. We have hierarchically divided the LLM types into four major categories (see Fig. 1).

Modern language model called ChatGPT was created by OpenAI. It is based on the GPT-3.5 architecture and was trained using a sizable amount of internet-sourced text data, including books, articles, wikis and websites II. ChatGPT is exceptional at producing human-like responses and having conversations with users. In computer vision (CV), researchers are actively engaged in the development of vision-language models inspired by the capabilities of ChatGPT. These models

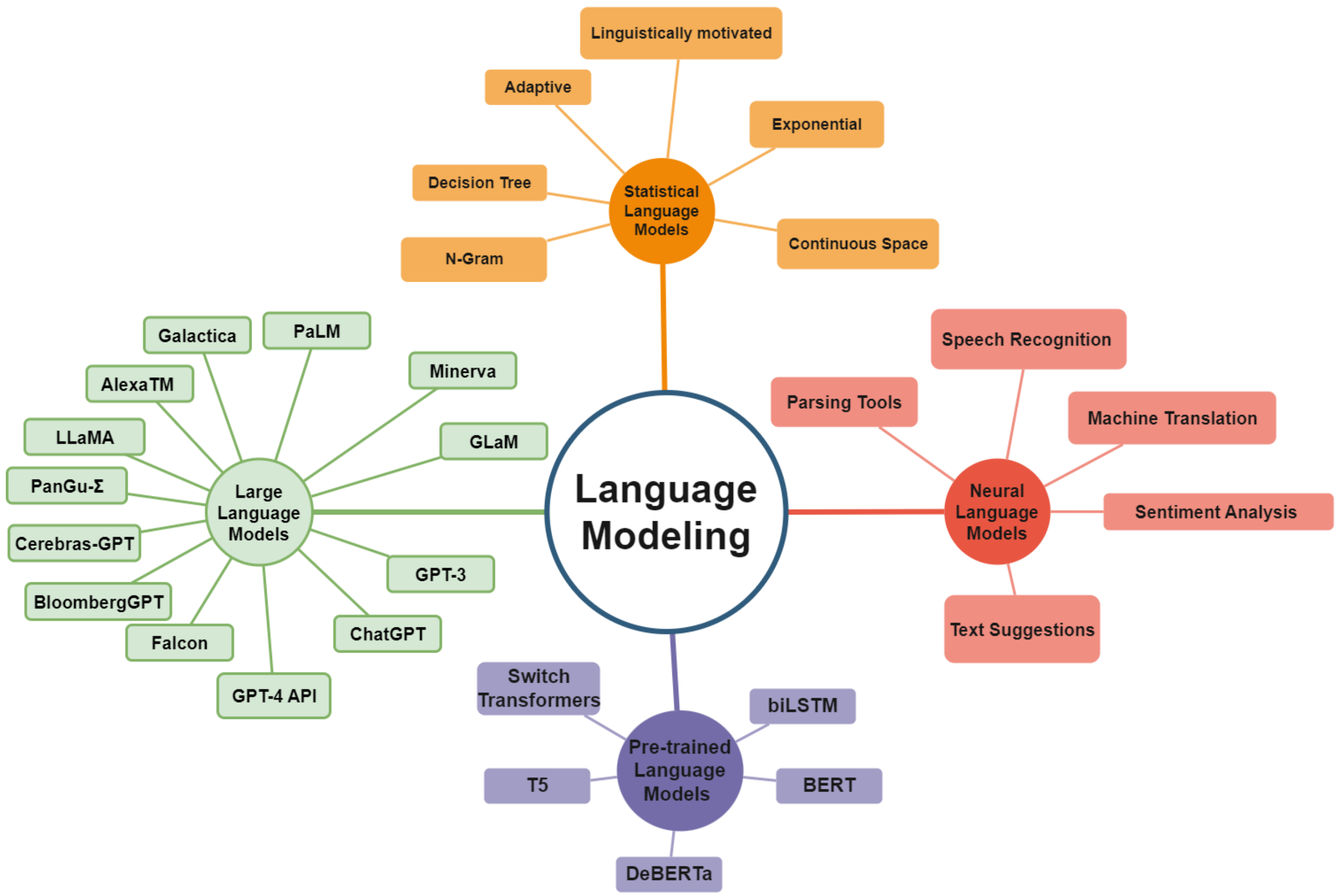


Fig. 1: Types of language modeling.

are specifically designed to enhance multimodal dialogues, where both visual and textual information are important [20]. Moreover, the advancements in the field have led to the introduction of GPT-4 [21], which has further expanded the capabilities of language models by seamlessly integrating visual information as part of the input. This integration of visual data empowers the model to effectively understand and generate responses that incorporate both textual and visual cues, enabling more contextually rich and nuanced conversations in multimodal settings.

A. Survey Motivation

The revolutionary ChatGPT has captivated the attention of the community, sparking a wealth of fascinating reviews and discussions on the advancements of LLMs and artificial intelligence [3], [22], [23], [24]. For example, the role of ChatGPT in education is evaluated in [25], healthcare in [26], finance in [27], on code writing capabilities in [28], impact on labour market in [29], legal aspects in [30], and an opinion paper in [5]. A comprehensive survey on LLMs presents advancements in pre-training, fine-tuning, utilization and capability evaluation of LLMs [3]. The recent progress in visio-language pre-trained models is discussed in [22]. The paper presents an overview of various techniques for encoding raw images and texts into single-modal embeddings as a fundamental aspect, and prevalent architectures of Vision-Language Pre-trained Models (VL-PTMs), focusing on their ability to effectively model the interaction between text and image representations.

Despite the growing number of studies on LLMs, there remains a scarcity of research focusing on their technical intricacies and effective utilization. In this review and tutorial article, our primary objective is to explore, learn, and evaluate language models across various domains. We delve into the working principles of language models, analyze different architectures of the GPT family, and discuss strategies for their optimal utilization. Furthermore, we provide detailed insights into writing prompts, and visual prompting techniques, leveraging GPT-plug-ins, and harnessing other AI/LLM tools. Our comprehensive examination also encompasses a discussion on the limitations associated with LLMs, including considerations related to security, ethics, economy, and the environment. In addition, we present a set of guidelines to steer future research and development in the effective use of LLMs. We hope that this research will contribute to a better understanding and utilization of LLMs. A list of commonly used acronyms in this article with definitions is given in Table I).

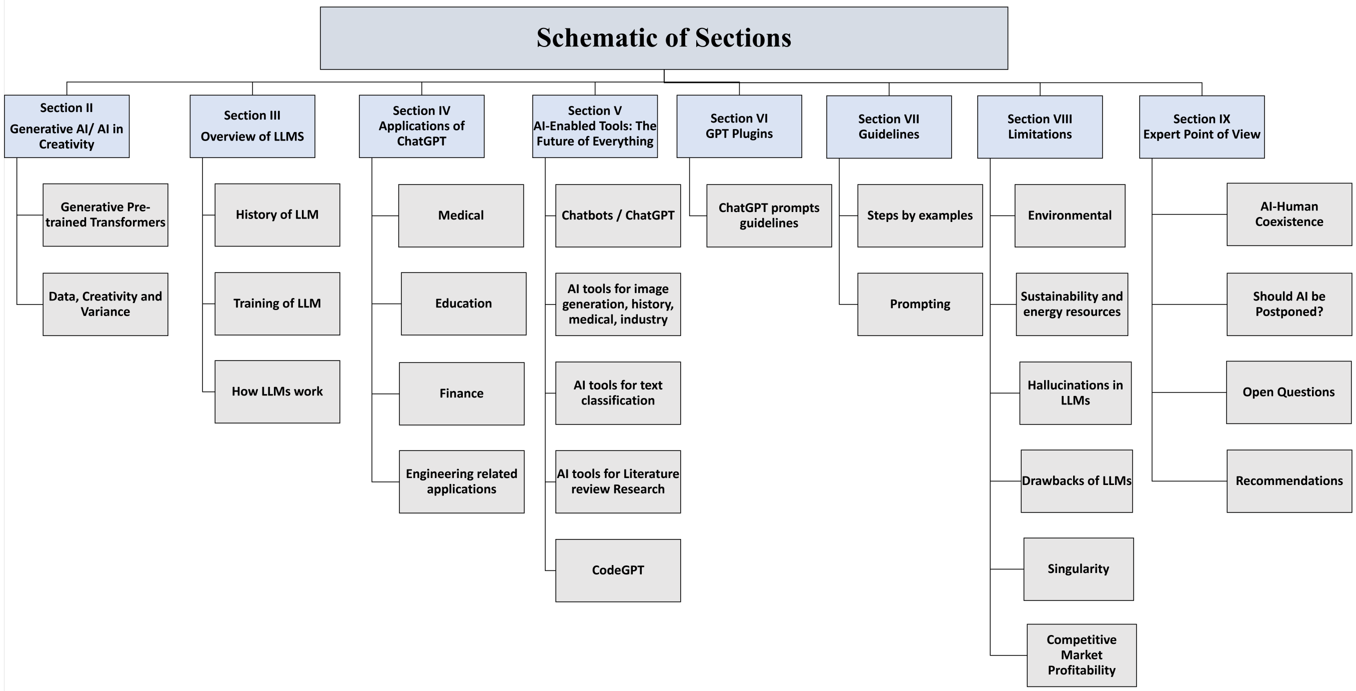


Fig. 2: Schematic of the overview of the survey at a glance.

B. Contributions

The main contributions of this article include:

1) Providing a comprehensive overview of LLMs, including their technical details, advancements, challenges, capabilities and limitations.
2) Presenting a state-of-the-art analysis and comparison of different LLMs.
3) Addressing ethical concerns about LLMs, including their computational requirements and potential for perpetuating biases.
4) Offering insights into the future potential of LLMs and their impact on society and demonstrating the applications of LLM through four practical use cases in the fields of medicine, education, finance, and engineering.
5) The paper is written in a **unique way to promote** practical usage of LLMs. Most of the content, including, images, figures and tables is generated using LLMs.

The survey paper is organized into the following sections. Section II provides an introduction to the role of AI in creativity, specifically focusing on generative pre-trained transformers and their significance. Section III presents an overview of LLMs, summarizing a brief history of LLMs and discussing their training and functionality. Section IV demonstrates the applications of LLM through four use cases in the fields of medicine, education, finance, and engineering. Section V explores AI-enabled tools that are expected to shape the future. Section VI discusses the practical use case of GPT plugins and their potential to enhance user productivity and efficiency. Section VII presents guidelines and working examples using prompting techniques. Section VIII proposes the limitations and drawbacks of the current state-of-the-art LLM. Section IX-D presents the impact of LLM on humans and society. Section XI presents expert opinions on the subject matter and the author's perspective on open unanswered avenues. Section XI concludes the survey paper. The overall structure and sections for better visibility are shown in the form of a schematic in Fig. 2.

TABLE II: **Pre-training data**. Mixtures of data used for pre-training LLaMA [31].

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3TB
C4	15.0%	1.06	783GB
Github	4.5%	0.64	328GB
Wikipedia	4.5%	2.45	83GB
Books	4.5%	2.23	85GB
ArXiv	2.5%	1.06	92GB
StackExchange	2.0%	1.03	78GB

II. Generative AI / AI in Creativity

Generative AI refers to AI systems primarily designed to generate content (text, images, audio and videos). It sets apart from AI systems with different functions, like classifying data (e.g., labeling images), grouping data (e.g., identifying customer segments with similar purchasing behavior), or making decisions (e.g., guiding an autonomous vehicle). Some common examples of generative AI systems are image generators (Midjourney or stable diffusion), Chatbots (ChatGPT, Bard, Palm), code generators (CodeX, Co-Pilot [40]) and audio generators (VALL-E).

Generative AI works by leveraging complex algorithms and statistical models to generate new content that mimics the patterns and characteristics of the training data. Generative AI systems can employ different techniques like Variational Autoencoders (VAEs) [41], Generative Adversarial Networks

TABLE III: Comparison of LLMs' Reasoning Performance. Notations: MMLU [32]: high school and college knowledge, GSM8K: elementary school math, MATH: very hard math and natural science. All current models struggle, BBH [33]: a collection of 27 hard reasoning problems, HumanEval [34]: a classical dataset for evaluating coding capability, C-Eval [35]: a collection of 52 disciplines of knowledge test in Chinese, TheoremQA [36]: a question-answering dataset driven by STEM theorems. [37], [32], [31], [38], [39], [35]

Model	Param.	Type	GSM8K	MATH	MMLU	BBH	HumanEval	C-Eval	TheoremQA
GPT-4	-	RLHF	92.0	42.5	86.4	-	67.0	68.7*	43.4
claude-v1.3	-	RLHF	81.8*	-	74.8*	67.3*	-	54.2*	24.9
PaLM-2	-	Base	80.7	34.3	78.3	78.1	-	-	31.8
GPT-3.5-turbo	-	RLHF	74.9*	-	67.3*	70.1*	48.1	54.4*	30.2
claude-instant	-	RLHF	70.8*	-	-	66.9*	-	45.9*	23.6
text-davinci-003	-	RLHF	-	-	64.6	70.7	-	-	22.8
code-davinci-002	-	Base	66.6	19.1	64.5	73.7	47.0	-	-
text-davinci-002	-	SIFT	55.4	-	60.0	67.2	-	-	16.6
Minerva	540B	SIFT	58.8	33.6	-	-	-	-	-
Flan-PaLM	540B	SIFT	-	-	70.9	66.3	-	-	-
Flan-U-PaLM	540B	SIFT	-	-	69.8	64.9	-	-	-
PaLM	540B	Base	56.9	8.8	62.9	62.0	26.2	-	-
LLaMA	65B	Base	50.9	10.6	63.4	-	23.7	38.8*	-
PaLM	64B	Base	52.4	4.4	49.0	42.3	-	-	-
LLaMA	33B	Base	35.6	7.1	57.8	-	21.7	-	-
InstructCodeT5+	16B	SIFT	-	-	-	-	35.0	-	11.6
StarCoder	15B	Base	8.4	15.1	33.9	-	33.6	-	12.2
Vicuna	13B	SIFT	-	-	-	-	-	-	12.9
LLaMA	13B	Base	17.8	3.9	46.9	-	15.8	-	-
Flan-T5	11B	SIFT	16.1*	-	48.6	41.4	-	-	-
Alpaca	7B	SIFT	-	-	-	-	-	-	13.5
LLaMA	7B	Base	11.0	2.9	35.1	-	10.5	-	-
Flan-T5	3B	SIFT	13.5*	-	45.5	35.2	-	-	-

(GANs) [42], or autoregressive models [43] to achieve the desired generation capabilities. These techniques allow the model to capture the underlying data distribution and generate new content that is similar to the training data, as shown in Fig. 3.

LLMs, such as ChatGPT, are a type of generative AI that is specifically designed to generate human-like language in response to a given prompt. These models are trained on massive amounts of textual data (see Table II), using techniques such as unsupervised learning to learn the statistical patterns of language. However, many people accord the capabilities provided by GPT models to "more data and computing power" instead of "better ML research".

A. Generative Pre-trained Transformers - GPT 3.5

The transformer-based architecture is a type of neural network that is well-suited for NLP tasks. It uses a stack of self-attention layers to learn long-range dependencies in text. Self-attention [44] is a mechanism that allows the model to learn the importance of each word in the input sequence, regardless of its position. This is important for NLP tasks, such as translation and question answering, where the meaning of a sentence can depend on the words that are far apart.

The architecture of GPT 3.5 is shown in Fig. 4, which consists of six major steps.

1) Input Encoding: The input to the GPT model is a sequence of tokens representing the data. Each token is converted into a high-dimensional vector representation through an embedding layer.
2) Transformer Encoder: The GPT model consists of multiple layers of transformer encoders. Each encoder layer has a self-attention mechanism and feed-forward neural networks. Self-attention allows the model to capture dependencies between different words in the input sequence, while the feed-forward networks process and transform the representations.
3) Contextual Embeddings: As the input sequence passes

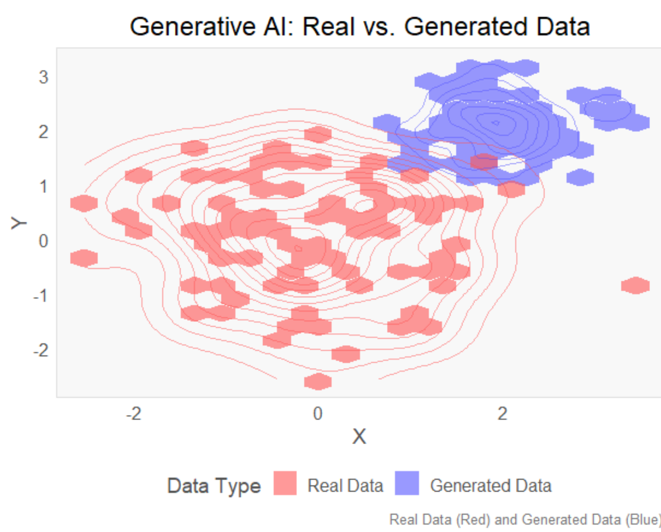


Fig. 3: A multivariate normal distribution applied to estimate the mean vector and covariance matrix of the underlying distribution using maximum likelihood estimation. Finally, we generate new synthetic data from the learned probability distribution. This demonstrates the concept of generative AI using probability distributions, where the AI learns to generate new data that is similar to the original data by modeling the underlying probability distribution

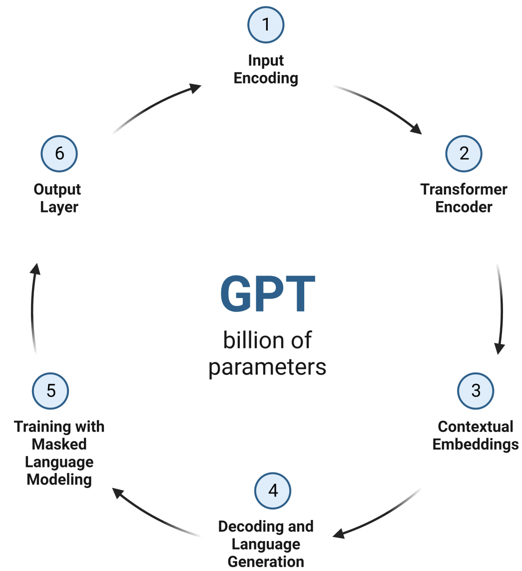


Fig. 4: The architecture of Generative Pre-trained Transformers

through the transformer encoders, each token's representation is updated in a contextualized manner. This means that the representation of each token is influenced by the surrounding tokens and their contextual information.

4) Decoding and Language Generation: Once the input sequence has been encoded through the transformer encoders, the GPT model can generate new text by predicting the probability distribution of the next token given the context. This process is typically done using a softmax activation function, which produces a probability distribution over all possible tokens in the vocabulary.
5) Training with Masked Language Modeling: During the pre-training phase, GPT models often employ a technique called Masked Language Modeling (MLM) [45]. MLM involves randomly masking certain tokens in the input sequence and training the model to predict those masked tokens based on the context. This helps the model learn contextual relationships and improves its ability to generate coherent text.
6) The 6th layer represents the output of the model.

LLMs have captured significant interest in recent years due to their remarkable performance across an extensive array of NLP tasks, including text generation, translation, summarization, question-answering, and sentiment analysis [46]. Constructed upon the foundation of the transformer architecture [14], these models exhibit an extraordinary capacity to process and generate human-like text by leveraging massive volumes of training data.

B. Data, Creativity, and Variance

As discussed previously, LLMs are developed by training large deep neural networks using text data made up of multiple different sources including but not limited to books, social media and also text from the internet such as poetry, songs, news articles etc. This diversity in the training mix allows the model to provide output text that is coherent just as a human written text may read. However, it should be noted that the "creativity" exhibited in LLMs goes beyond the regurgitation of data that it may have seen during the training process. To produce text creatively, the deep learning model of the LLM needs to form an understanding of the text used in its training in aspects related to language, tone and writing patterns etc. This way the LLM is able to generate responses to user queries that are creative and genuine in terms of language writing by combining the different types of input information it ingested during training together to generate meaningful results for the provided query.

A fundamental criterion for gaining the capability of this creativity is to have sufficient variance, which indicates to the model's ability to produce an unexpected output. In short, variance ensures that there is randomness in a model's output, and it is introduced to enable it to generate sufficiently good results over a range of output results. By introducing variance in the model, one can increase the diversity of output content generated which goes beyond the scope of what the training data consisted of.

It is acknowledged that since the release and mainstreaming of LLMs by users of all walks of life, some have complained of LLM getting stuck in a cycle of similar answers, especially if a complex query has been asked of it multiple times. For e.g., Microsoft found that its Bing AI powered by ChatGPT tends to get repetitive in its responses after 15 consecutive chats [1]. While this problem has mitigated since then by taking measures such as refreshing context and/or introducing limits to the questions asked per session, this does question the variance capability of LLMs. There is a philosophical consideration here, LLMs have been associated with being "God-like AI" due to their exhibited creativity in different fields. From a human scenario, an example could be considered fingerprints, each one of the more than eight billion people in the world has unique fingerprints which can be used to identify them. However, from a philosophical perspective, it could be of interest how the size of the human population affects the variance of this space of potential of different fingerprints for humans? Therefore, it is important to appreciate the total "space" of creativity that LLMs have by design and consider creativity in that respect.

III. Overview of LLMs

LLMs have revolutionized the field of artifical intelligence and have found applications in various domains, including communication, content generation, and knowledge dissemination. In this Section, we briefly discuss the history, training, and working of LLMs.

A. History of LLM

LLMs are a type of AI model that can process and generate natural language text. These models are typically trained on

[1]https://www.zdnet.com/article/long-chats-confuse-bing-chat-so-microsofts-chatgpt-powered-bot-is-getting-new-limits/

massive amounts of text data and use deep learning techniques to learn the patterns and structures of language [47]. The history of LLMs can be traced back to the early days of NLP research [6].

The first language models were developed in the 1950s and 1960s. These models were rule-based and relied on hand-crafted linguistic rules and features to process language. They were limited in their capabilities and were not able to handle the complexity of NLP [48].

In the 1980s and 1990s, statistical language models were developed. These models used probabilistic methods to estimate the likelihood of a sequence of words in a given context. They were able to handle larger amounts of data and were more accurate than rule-based models [49]. However, they still had limitations in their ability to understand the semantics and context of language [50].

The next major breakthrough in language modeling came in the mid-2010s with the development of neural language models [51]. These models used deep learning techniques to learn the patterns and structures of language from large amounts of text data. The first neural language model was the recurrent neural network language model (RNNLM), which was developed in 2010. RNNLM was able to model the context of words and produce more natural-sounding text than previous models [52].

In 2015, Google introduced the first large-scale neural language model called the Google Neural Machine Translation (GNMT) system [53]. This model was trained on massive amounts of bilingual text data and was able to achieve state-of-the-art performance on machine translation tasks. Figure 5 shows the evolution of LLMs.

The development of LLMs continued with the introduction of the Transformer model in 2017 [14]. The Transformer was able to learn the longer-term dependencies in language and allowed for parallel training on multiple Graphical Processing Units (GPUs), making it possible to train much larger models [54].

The release of OpenAI's GPT-1 [55] in 2018, marked a significant advance in NLP with its transformer-based architecture. With 117 million parameters, GPT-1 could generate contextually relevant sentences, demonstrating the potential of transformers in revolutionizing NLP tasks [56]. It was trained using a two-step process of unsupervised pre-training and supervised fine-tuning, a methodology that generated considerable interest in the academic and research community. Although GPT-1 had its limitations, it set the stage for subsequent, more powerful models, propelling a new era of AI research and highly-competitive research in LLMs (see Fig. 5).

In 2020, OpenAI released the largest language model to date, GPT-3, which was trained on a massive amount of text data and was able to generate highly coherent and natural-sounding text [57]. GPT-3 demonstrated the potential of LLMs for a wide range of NLP tasks [58].

Inspired by the success of GPT-3, OpenAI recently announced and began working on the development of the next iteration of their language model, GPT-4 [59]. GPT-4 is expected to be even larger and more powerful than GPT-3, with the ability to generate even more coherent and natural-sounding text.

While LLMs have found use in a number of different applications, they can also be biased and produce inaccurate or harmful outputs. There are a number of factors that can contribute to the weakness of an LLM. One factor is the size of the training dataset. If the dataset is too small, the model may not be able to learn to generalize to new situations. Another factor is the quality of the training dataset. If the dataset contains biased or inaccurate information, the model will learn to reflect that bias or inaccuracy.

There are also a number of factors that can contribute to the strength of an LLM. One factor is the size of the model. Larger models have more parameters, which allows them to learn more complex relationships between words and concepts. Another factor is the architecture of the model. Some architectures are better suited for certain tasks than others.

In order to evaluate the performance of an LLM, it is important to consider the following factors:

- The size of the training dataset.
- The quality of the training dataset.
- Parametric size of the model
- Complexity of the architecture of the model.
- The task that the model is being evaluated on.

It is also important to note that LLMs are still under development, and their performance can vary depending on the specific task and the environment in which they are used. The following subsection discusses the most common LLMs, their training and working principles.

B. Training of LLMs

Training large language models involves several key steps that are fundamental to their successful development. The process typically begins with the collection and preprocessing of a massive amount of text data from diverse sources, such as books, articles, websites, and other textual corpora (see Table. IV). The curated dataset [61] serves as the foundation for training the LLMs. The training process itself revolves around a technique known as unsupervised learning, where the model learns to predict the next word in a sequence given the preceding context. This task is commonly referred to as language modeling. LLMs utilize sophisticated neural network architectures, such as Transformers, which enable them to capture complex patterns and dependencies in language. The training objective is to optimize the model's parameters to maximize the likelihood of generating the correct next word in a given context [3]. This optimization is typically achieved through an algorithm called stochastic gradient descent (SGD) or its variants, combined with backpropagation, which computes gradients to update the model's parameters iteratively.

- Generative Pre-trained Transformer 3 (GPT-3): GPT-3 is one of the most advanced and largest language models developed by OpenAI [57]. It represents a significant breakthrough in NLP and has garnered considerable attention due to its impressive capabilities [3]. GPT-3 follows a transformer-based architecture, which allows it to capture complex linguistic patterns and dependencies

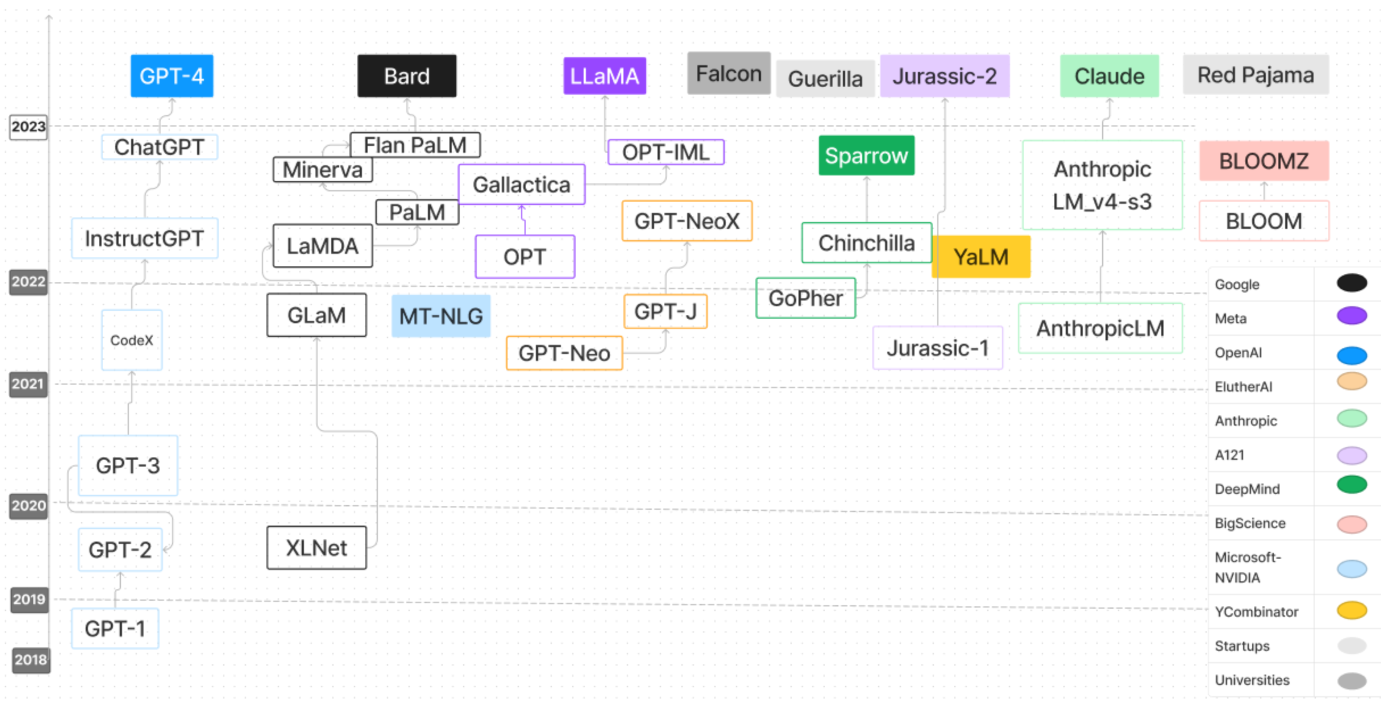


Fig. 5: Evolution of LLMs across different research and commercial organizations.

TABLE IV: State-of-the-art for LLM training pipeline [60]. Notations: RM: Reward Modeling, RL: Reinforcement Learning, SFT: Supervised Fine-tuned.

Stage	Pretraining	Supervised-Finetuning	Reward Modeling	Reinforcement Learning
Dataset	Raw Internet II	Demonstration	Comparisons	Prompts
Algorithm	Language Modeling	Language Modeling	Binary Classification	Reinforcement Learning
Model	Base Model	SFT Model	RM Model	RL Model
Resources	100s of GPUs months of training deployable	1-100 of GPUs days of training deployable	1-100 of GPUs days of training not deployable	1-100 of GPUs days of training deployable

in text [62]. The model consists of a stack of transformer layers, enabling it to process and generate text at various levels of abstraction. With a staggering number of approximately 175 billion parameters, GPT-3 is currently one of the largest language models ever created. The training process of GPT-3 involves unsupervised learning on a massive corpus of publicly available text data from the internet. By leveraging its enormous size and extensive training data, GPT-3 has acquired a broad understanding of language and can generate human-like text across various topics [63].

- Bidirectional Encoder Representations from Transformer (BERT): BERT is a prominent language model with significantly advanced NLP tasks. Its training process comprises pretraining and fine-tuning stages [64]. During pretraining, BERT learns a general language representation from large-scale unlabeled text data. It employs masked language modeling (MLM) and next-sentence prediction (NSP) tasks. MLM involves masking a portion of input tokens and training the model to predict the original masked tokens, fostering bidirectional context understanding [65]. NSP trains BERT to predict whether a second sentence follows the first, enhancing coherence comprehension. After pretraining, BERT undergoes fine-tuning on specific tasks with labeled data. Fine-tuning tailors BERT's learned representations to target tasks, such as sentiment analysis or named entity recognition. It employs backpropagation and gradient descent optimization to update model parameters. Training BERT demands significant computational resources [56], utilizing high-performance hardware like GPUs or Tensor Processing Units (TPUs) or field programmable gate arrays (FPGAs) [66], [67], [68]. Techniques such as layer normalization, residual connections, and attention mechanisms inherent in the transformer architecture further enhance BERT's capacity to capture intricate dependencies and long-range contextual relationships.
- eXtreme Language understanding Network (XLNet): XLNet is a generalized autoregressive pre-training method that surpasses the limitations of traditional left-to-right or right-to-left language modeling. XLNet is trained using a permutation-based approach that differs from traditional autoregressive models [69]. In the training process, rather than predicting the next word given the

previous words in a fixed order, XLNet considers all possible permutations of the input sequence and models the probability of each permutation. This allows XLNet to capture dependencies in both directions, thus addressing the limitations of sequential left-to-right or right-to-left modeling [70]. The training of XLNet involves two key steps: unsupervised pretraining and supervised fine-tuning. During unsupervised pretraining, XLNet learns to predict words conditioned on the entire input context by maximizing the expected log-likelihood over all possible permutations. This is achieved using a variant of the transformer architecture, similar to models like BERT. The permutation-based objective function used in XLNet training presents unique challenges. Unlike traditional autoregressive models that can rely on the causal order of words for prediction, XLNet needs to consider all possible permutations, resulting in an exponentially large number of training instances. This makes the training process computationally intensive and requires efficient strategies, such as ”factorized sampling,” to sample a subset of permutations during each training iteration. Another difficulty in training XLNet is the need for large-scale computing resources [3], [71], [72]. The vast number of possible permutations and the large model size contribute to increased memory and computation requirements. Training XLNet often necessitates distributed training on multiple GPUs or TPUs and can take significant time [3].

- Text-to-Text Transfer Transformer (T5): T5, developed by Google, is a versatile language model that is trained in a ”text-to-text” framework. The training process of T5 involves two main steps: pretraining and fine-tuning. During pretraining, T5 is trained on a massive corpus of publicly available text from the internet. The objective is to learn a generalized representation of language that can be applied to a wide range of tasks.The key innovation of T5 is the formulation of all tasks as text generation problems. This means that every task, including text classification, summarization, translation, and question answering, is cast into a text-to-text format. For example, instead of training T5 to answer questions directly, it is trained to generate the complete answer given the question and relevant context. In the pretraining phase, T5 is trained using a variant of the transformer architecture. The transformer model allows T5 to capture long-range dependencies and effectively model the contextual relationships in the input text [73]. The pretraining objective is typically based on maximum likelihood estimation, where T5 is trained to predict the target text given the source text. Once pretraining is complete, T5 undergoes fine-tuning on specific downstream tasks [73].
  One of the challenges in training T5 is the availability of large-scale labeled datasets for various tasks. Fine-tuning requires task-specific labeled data, and the quality and quantity of the data play a crucial role in the model's performance [3]. Additionally, the computational resources required to train T5 can be substantial, as the model is computationally intensive due to its transformer architecture and the size of the pre-trained parameters.
- Conditional Transformer Language Model (CTRL): CTRL is a language model designed to generate text based on specific control codes or prompts. It is trained using a two-step process: pretraining and fine-tuning. During pretraining, CTRL is trained on a large corpus of publicly available text data [74]. The objective of pre-training is to teach the model to understand and generate coherent text based on different control codes or prompts [75]. The training data includes diverse sources such as books, articles, websites, and other text documents. The training process involves utilizing the transformer architecture, similar to models like BERT and GPT.
  The model is trained to predict the next word or phrase in a given context, learning the statistical patterns and linguistic structures of the language. One of the unique aspects of CTRL is its conditioning of control codes or prompts. These control codes guide the model's text generation process, allowing users to specify the desired style, topic, or other characteristics of the generated text. The control codes act as explicit instructions to guide the model's behavior during both training and inference. The fine-tuning phase of CTRL is crucial for adapting the model to specific tasks or domains. Fine-tuning involves training the pre-trained CTRL model on task-specific datasets with control codes. The model is exposed to task-specific prompts and is trained to generate text that aligns with the desired output or behavior for the given task.

C. How LLMs work

At their core, LLMs, are a type of AI that can mimic human intelligence. They function by employing advanced statistical models and deep learning techniques to process and understand extensive amounts of text data [76]. These models learn the intricate patterns and relationships present in the data, enabling them to generate new content that closely resembles the style and characteristics of a specific author or genre [77].

The process begins with pre-training, during which the LLM is exposed to a massive corpus of text from various sources such as books, articles, and websites. Through unsupervised learning, the model learns to predict the next word in a sentence based on the context of the preceding words. This allows the model to develop an understanding of grammar, syntax, and semantic relationships [74]. As shown in the LLMs pre-training pipeline in Figure 6, the first step is pre-training corpus sources which can be roughly divided into two categories: general data and specialized data. Following the collection of a huge amount of text data, it is critical to preprocess the data in order to generate the pre-training corpus, particularly by removing noisy, redundant, unnecessary, and potentially poisonous material [78] [79].The second stage involves quality filtering to remove the low quality and unwanted data from the training corpus using some techniques such as the language filtering, statistic filtering and keyword filtering [79]. Third stage is deduplication, where previous research [80] discovered that duplicate data in a corpus reduces the diversity of LMs, causing the training process to become

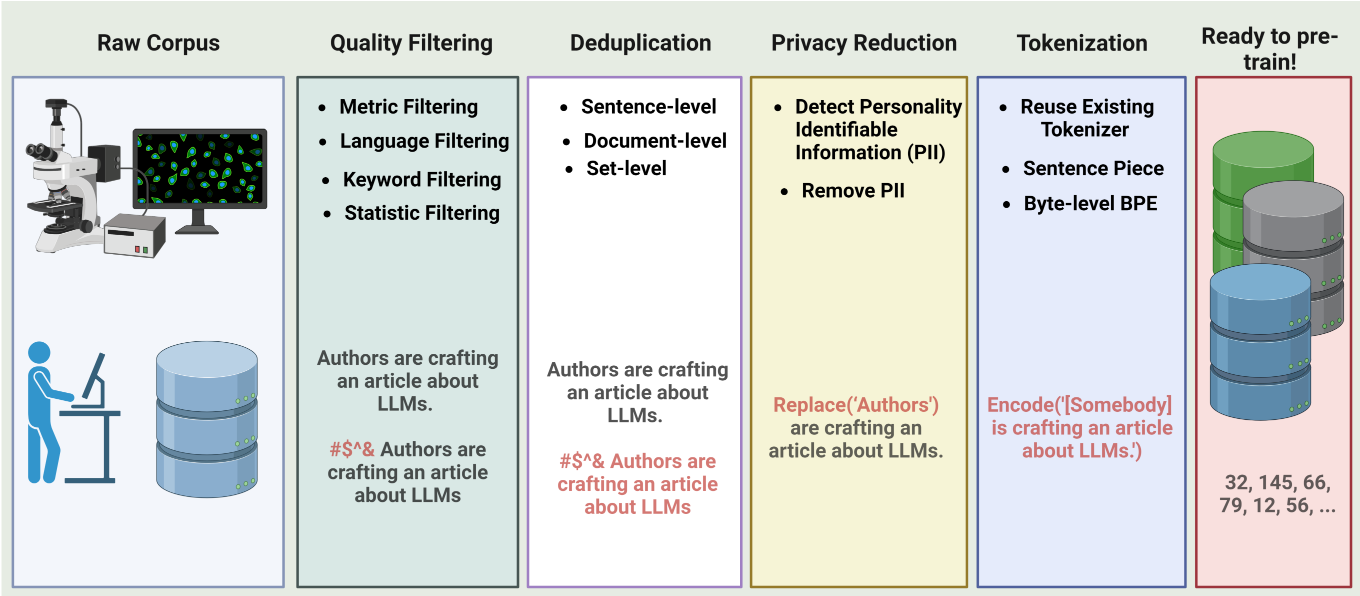


Fig. 6: The illustration of a data processing pipeline for pre-training LLMs.

unstable and thereby affecting model performance [80]. The fourth stage privacy reduction, it is crucial to address privacy concerns related to the use of web-based data for pre-training language models. This data often includes user-generated content containing sensitive or personal information, thereby posing a potential risk of privacy breaches [81]. Therefore, it is essential to undertake privacy redaction measures to remove personally identifiable information (PII) from the pre-training corpus. The last step is tokenization. It is an important step in data preprocessing as well. It seeks to segment raw text into sequences of individual tokens, which are then fed into LLMs.

Following pre-training, the LLM undergoes fine-tuning, which involves training the model on a specific task or domain. During this phase, the model is provided with labeled examples and guided to generate more accurate and contextually appropriate responses for the target task [74]. Fine-tuning allows the LLM to specialize in various applications, such as language translation, question-answering, or text generation.

The success of LLMs lies in their ability to capture the statistical patterns and linguistic nuances present in the training data [75]. By processing and analyzing vast amounts of text, LLMs gain a comprehensive understanding of language and are able to generate coherent and contextually relevant responses. During the inference stage, when interacting with an LLM, a user inputs a prompt or query. The model processes the input and generates a response based on its learned knowledge and context. This response is generated using probabilistic methods that consider the likelihood of various words or phrases given the input context. In Figure 6, we show the data pre-processing pipelines for pre-training LLMs.

IV. Applications of Large Language Models

Given LLMs wide range of applications, in this section, we provide a discussion of their use in the fields of medicine, education, finance, and engineering. The selection of medical, education, finance, and engineering as the applications for LLM is based on their significance, relevance, and potential impact within their respective domains. These applications demonstrate the versatility and potential of LLMs in addressing complex challenges and supporting human endeavors.

A. Medical

LLMs like ChatGPT have exhibited remarkable potential in diverse healthcare applications, particularly in the field of medicine. They have been successfully employed in medical education, radiologic decision-making, clinical genetics, and patient care, as supported by several studies [82],[83]. In medical education, ChatGPT has emerged as an interactive tool that aids learning and problem-solving [84]. Notably, ChatGPT's performance in the United States Medical Licensing Exam (USMLE) was comparable to or exceeded the passing threshold, indicating its proficiency in medical knowledge without requiring specialized training or reinforcement [84]. Moreover, ChatGPT's explanations displayed a high level of concordance and insightful understanding [82].

According to a study conducted by Rao et al. [85], it is anticipated that specialized AI-based clinical decision-making tools will emerge in the future. This study emphasizes the potential of ChatGPT in radiologic decision-making, highlighting its feasibility and potential benefits in improving clinical workflow and ensuring responsible use of radiology services. Similarly, Kung et al. [82] concluded in their research that LLMs, including ChatGPT, have the capacity to enhance the delivery of individualized, compassionate, and scalable healthcare. These models can assist in medical education and potentially aid in clinical decision-making.

In the domain of clinical genetics, a paper by Duong and Solomon [86] found that ChatGPT's performance did not significantly differ from humans when answering genetics-related questions. However, the model demonstrated better accuracy on memorization-type questions compared to questions requiring critical thinking. Notably, this study also highlighted that ChatGPT provided varying answers when asked the same question multiple times, providing plausible explanations for both correct and incorrect responses. Furthermore, Fijacko [87] conducted a study to evaluate ChatGPT's accuracy in answering questions related to life support and resuscitation. The findings revealed that ChatGPT demonstrated the ability to provide accurate answers to a majority of the questions on the American Heart Association's Basic Life Support and Advanced Cardiovascular Life Support exams.

In the field of neurosurgical research and patient care, ChatGPT has been investigated for its potential role in various aspects, including gathering patient data, administering surveys or questionnaires, and providing information about care and treatment [88]. Nevertheless, the implementation of such technologies necessitates careful consideration to ensure their effectiveness and safety. The integration of biotechnology and AI to address global challenges and advance sustainable development goals is examined in a research paper that encompasses a wide range of AI applications in the life sciences. These applications encompass decision support, NLP, data mining, and machine learning. The authors underscore the significance of reproducibility in the development of AI models and highlight ongoing research issues and challenges in these domains [89]. Furthermore, AI-powered chatbots like ChatGPT hold the potential to enhance patient outcomes by facilitating communication between patients and healthcare professionals. Leveraging NLP, these chatbots can provide patients with information about their care and treatment in a more accessible manner [90]. A database for Covid-19 drug repurposing using NLP is proposed in [91]. There have several tools already in use that allows the system to interact with patients such as Ada Health, Babylon Health and Buoy Health. The recent popularity of LLMs can potentially not only improve patient confidence in interacting with such chatbots but also improve upon the services provided. In fact, there are tools developed to assist medical practitioners. One such tool is XrayGPT [92], it can be used for automated analysis of X-ray images and have the user/patient ask questions about the analysis. Through the chats, the user can get insight into their condition through an interactive chat dialogue.

B. Education

The impact of AI on education has been a topic of much discussion in recent years. One area where AI is having a significant impact is in the realm of student assignments and exams. Since the advent of ChatGPT developed by OpenAI, the way students interact with educational materials, assignments and coursework has become different [93] [94] [95]. One of the main advantages of using ChatGPT and AI bots in education is that they can help students complete their assignments more efficiently [96]. ChatGPT is capable of generating high-quality responses to a wide range of prompts, which can save students time and effort when they are working on assignments. Additionally, AI bots can help to automate the grading process, which can reduce the workload for teachers and enable them to provide more detailed feedback to students.

Another advantage of using ChatGPT and AI bots in education is that they can provide personalized learning experiences for students. AI bots can analyze a student's performance on previous assignments and exams and use this data to generate personalized recommendations for future work. This can help students to identify their strengths and weaknesses and focus their efforts on areas where they need to improve. Khan Academy, a nonprofit educational organization, has shown interest in utilizing ChatGPT for its business. They have developed an AI chatbot called Khanmigo, which serves as a virtual tutor and classroom assistant [97]. The goal of incorporating ChatGPT into their platform is to enhance tutoring and coaching experiences by providing one-on-one interactions with students. This is optimism about the potential of AI, including ChatGPT, in the field of education. The use of AI in assisting with tutoring and teaching also dispels the notion that its primary use is for cheating. Undoubtedly, AI technology is still in its early stages, but it holds promise for helping students and addressing their diverse needs [98].

However, there are also some potential drawbacks to using ChatGPT and AI bots in education. One concern is that these technologies may lead to a loss of creativity and critical thinking skills among students. If students rely too heavily on AI bots to complete their assignments and exams, they may not be developing the skills necessary to think critically and solve problems on their own [96].

1) Learning in the age of AI: Another major assistance that these bots such as ChatGPT can offer is the provision of assistance in designing a course in an academic setting. AI chatbots can serve as a valuable tool to aid in various aspects of syllabus preparation. Course objectives can be generated, relevant topics identified, curricula structured, learning resources gathered and reviewed, assessment methods defined, engaging learning activities established, and a well-balanced course schedule created. The iterative process of interacting with ChatGPT enables refinement and enhancement of the syllabus based on the model's suggestions and insights. It is important to note that ChatGPT acts as a supportive tool, augmenting the expertise and input of experienced educators. The collaboration between human and AI in the course syllabus design process facilitates the development of comprehensive and effective learning plans that align with desired learning outcomes.

2) Major issues for AI in Education: One of the major concerns is the utilization of these tools without proper training. Even if LLMs can provide answers to a plethora of questions and assist the user in providing answers to questions on their fingers, it is important for pupils in educational institutions to be trained sufficiently to make the best use of LLMs capabilities.

Another concern is that the use of AI bots in education could lead to increased inequality [99]. Students who have access to these technologies may have an unfair advantage over those who do not, which could exacerbate existing inequalities in education. Additionally, the use of AI bots could lead to a decrease in the number of teaching jobs available, which could further widen the gap between those who have access to education and those who do not. In conclusion, the use

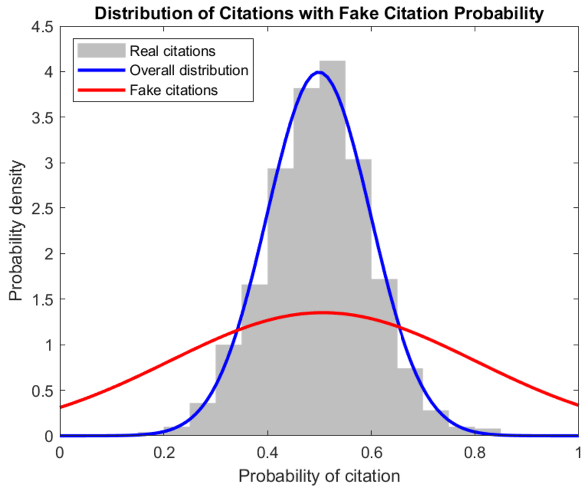


Fig. 7: The overall citation distribution is represented by a gray histogram with 50% opacity. The blue line represents the probability density function of the overall citation distribution, estimated using a Gaussian kernel density estimator. The red line represents the probability density function of fake citations, estimated using a Gaussian kernel density estimator. Prompt "SUPPOSE you have the probability distribution of a universe, and when you sample from them there is a high probability of fake citation, I want to show this concept on a single probability distribution" Please provide MATLAB code.

of ChatGPT and AI bots in education has both pros and cons. While these technologies can help students complete assignments more efficiently and provide personalized learning experiences, they may also lead to a loss of critical thinking skills and increased inequality. As AI continues to transform the field of education, it will be important to carefully consider these potential benefits and drawbacks and work to minimize the discussed negative consequences that may arise.

C. Finance

LLMs are making significant advancements in the finance industry with applications ranging from financial NLP tasks, risk assessment, algorithmic trading, market prediction and financial reporting. LLM's such as BloombergGPT[27], a 50 billion parameter large language model trained on large diversified financial corpus, has revolutionized financial NLP tasks such as news classification, entity recognition and question answering. By utilizing the huge amount of financial data available, it is able to enhance customer services drastically by efficiently handling customer queries and providing them with excellent financial advisory.

In addition, LLMs are being used for risk assessment and management, by analyzing past market trends and data, it is able to identify potential risks and provide mitigation steps through different financial algorithms. Financial institutions can use it for better decision making such as credit risk assessment, loan approvals and investments. Algorithmic Trading is another application that can leverage LLM models to identify potential opportunities in the trading market by using its predictive and analyzing capabilities.

However, due to the sensitivity of the financial information and privacy concerns, techniques like data encryption, redaction and data protection policies should be implemented so that these LLMs can be used efficiently in accordance with data protection policies. In this regard, a recent proposition suggested is FinGPT [100] which is an open-source LLM tailored for finace. It is expected that more work will be carried out in this space.

D. Engineering related applications

LLMs have gained substantial attention across various fields, and their potential applications in engineering domains are increasingly being explored. For instance, ChatGPT has diverse applications in software engineering, including code generation, debugging, software testing, NLP, documentation generation, and collaboration. It enables developers to generate code snippets, identify and fix errors, generate test cases, analyze user requirements, create user interfaces, generate software documentation, and facilitate collaboration within development teams. ChatGPT's language understanding and generation capabilities enhance efficiency, streamline workflows, and foster effective communication in software engineering.

In software engineering, ChatGPT can be employed to generate code snippets based on natural language descriptions of desired functionality. This feature saves developers time and improves overall efficiency, allowing them to focus on higher-level design aspects [101]. Additionally, ChatGPT can assist in debugging code by leveraging its language understanding capabilities to identify errors and suggest potential fixes, thereby streamlining the debugging process and reducing development time. The use of ChatGPT extends to software testing, where it can generate test cases and test data based on natural language descriptions of desired test scenarios. This approach enhances the efficiency and effectiveness of software testing, ensuring comprehensive coverage and accurate validation of the software's functionality.

In mechanical engineering, Tiro [102] has study the possibility of ChatGPT utilization to various calculations in mechanical engineering, Tiro encountered instances where incorrect procedures, formulas, or results were provided. None of the tasks yielded an exact solution, leading them to discontinue further research. Based on Tiro findings, it can be concluded that, at the current stage of AI development, ChatGPT should not be relied upon for solving engineering practice problems. Furthermore, caution should be exercised in using ChatGPT for such applications, as incorrect results can have potential consequences.

In Mathematics, some attempts have been done such as Wardat et al [103] found that ChatGPT holds potential for assisting in teaching mathematics by providing interactive and dynamic learning experiences. It can generate customized examples and problem-solving strategies tailored to individual student needs, fostering personalized learning. Moreover, it can serve as a virtual tutor, offering real-time feedback and guidance, identifying areas of difficulty, and suggesting alternative approaches. As an AI language model, ChatGPT is capable of performing mathematical calculations and solving math equations. However, the accuracy and effectiveness of ChatGPT solutions may depend on various factors such as the complexity of the equation, the accuracy of the input data, and the instructions given to ChatGPT. Frieder et al., [104] has investigated the mathematical capabilities of ChatGPT by testing it on publicly available datasets, as well as hand-crafted ones, and measuring its performance against other models trained on a mathematical corpus, such as Minerva. They also test whether ChatGPT can be a useful assistant to professional mathematicians by emulating various use cases that come up in the daily professional activities of mathematicians (question answering, theorem searching).

However, it is essential to acknowledge the limitations of ChatGPT, including the possibility of generating incorrect responses or failing to address complex mathematical concepts adequately. Therefore, it should be utilized as a supplemental tool alongside traditional teaching methods and human supervision to ensure accuracy and quality in teaching mathematics.

In manufacturing, Wang et al. [105] conducted an evaluation of ChatGPT's capabilities in supporting design, manufacturing, and engineering education tasks. The results indicate that ChatGPT is impressive in providing information, generating coherent and structured content, and proposing initial solutions. The authors recommended a technology development roadmap to successfully integrate ChatGPT into the manufacturing industry. Therefore, in manufacturing, ChatGPT struggles to understand questions and lacks the ability to properly use knowledge to generate correct solutions and it can even fabricate non-existing rules or equations in order to generate solutions. Similarly, Badini et al. [106], performed a study in additive manufacturing troubleshooting and evaluated ChatGPT's expertise in technical matters, focusing on the evaluation of printing parameters and bed detachment, warping, and stringing issues for Fused Filament Fabrication (FFF) methods using thermoplastic polyurethane polymer as feedstock material. It was found that ChatGPT provided remarkable accuracy, correctness, and organization in its responses and its approach to problem-solving offered valuable insights in addressing hurdles. In particular, for the specific technical issues of warping, bed detachment, and stringing, ChatGPT demonstrated its ability to provide hierarchical and logically organized responses while taking into account given information and constraints. Furthermore, it was also able to fine-tune printing parameters for different types of TPU filaments, showing its ability to relate the mechanical properties of the filament material to the printing parameters. Finally, the authors recommended integrating ChatGPT into an Additive Manufacturing software platform to provide real-time suggestions and optimization for users, which can enhance the efficiency and quality of the Additive Manufacturing process.

V. AI-Enabled Tools: The Future of Everything

AI tools are becoming increasingly powerful and versatile. They can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way. These powerful tools are designed to understand and generate human-like text, offering a wide range of applications and benefits. AI tools built upon LLMs provide developers and researchers with accessible APIs and libraries to leverage the capabilities of these models. They offer a user-friendly interface for tasks like text generation, image generation, coding, sentiment analysis, language understanding, and content recommendation [3]. In this Section, we discuss various AI-enabled tools based on LLMs.

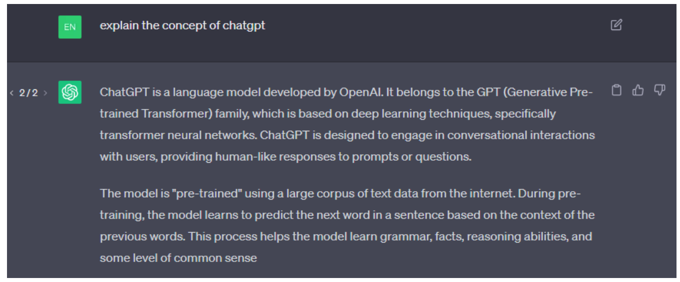


Fig. 8: A simple educational conversation with ChatGPT

A. *Chatbots / ChatGPT*

Chatbots are frequently used in customer service applications where they can respond to queries, offer assistance, and fix problems [107]. They can also be utilised for other purposes, including entertainment, healthcare, and education. Chatbots and LLMs are often used together to create more sophisticated and engaging conversational experiences. For example, a chatbot might use an LLM to generate text for its responses. Some of the popular chatbots include ChatGPT, Google Bard, and Microsft Bing. In Fig. 8, we show a simple educational conversation with ChatGPT.

Another

1) Comparison between Chatbots: ChatGPT and Google Bard are two of the most popular LLMs available today [108]. Both models are capable of generating text, translating languages, writing different kinds of creative content, and answering your questions in an informative way. However, there are some key differences between the two models, such as ChatGPT is more creative, while Google Bard is more authentic. Table VII presents a comparison between ChatGPT, Google Bard, and Microsoft Bing Chatbots.

B. *AI tools for image generation, history, medical, industry*

Table V showcases the output of image generation using various prompts. In total, nine different prompts were used, these required the AI model to generate humans and natural scenery. The first four prompts tended to depiction of famous personalities (sportsmen and politicians in this case), Muhammad Salah, Lionel Messi, Mike Tyson and Imran Khan. The prompts used were "Mo Salah playing cricket", "Lionel Messi playing tennis", "Mike Tyson playing football" and "Imran Khan as a hero". The second prompt used was regarding the famous painting Monalisa. The prompt was "Generate an image of Monalisa showing her teeth in a wedding ceremony". The third prompt related to natural scenery and was written as "Area of rocks, deep inside the forest, divine domain".

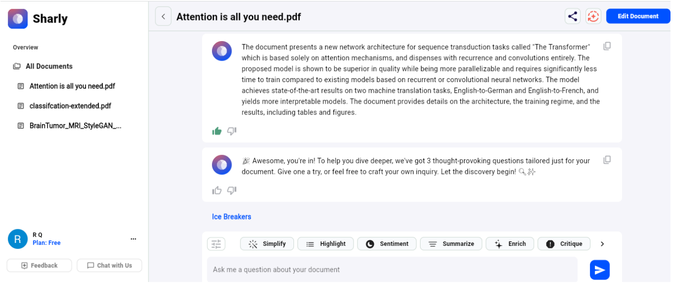


Fig. 9: An example of PdfGPT. Upload any PDF document and start chatting. It helps in summarizing, highlighting, critiquing, and simplifying the content.

Lastly, the fourth prompt also centered around the generation of humans. In this case, three prompts were given, ”A man kissing a girl”, ”Generate an image of a guy” and Generate and image of a woman”.

C. AI tools for text classification

AI tools are increasingly being used for text classification. Text classification is the process of assigning a category to a piece of text [109]. For example, a text classification tool could be used to classify emails as spam or not spam or to classify news articles as business, sports, or entertainment. Some of the popular libraries include Scikit-learn, NLTK, and Spacy.

D. AI tools for Literature review Research

AI tools are increasingly being used to assist with literature review research. These tools can be used to automate tasks such as: Identifying relevant literature, extracting information, and summarizing the content [110]. One such tool is PDFGPT [111], which uses the GPT-3 model to generate responses to user queries. PDFGPT can be used to extract information from PDF files, answer questions about the content of PDF files, and generate summaries of PDF files. An example of PDFChat is shown in Fig. 9.

Another interesting AI tool is elicit.org, which helps automate literature reviews. The website offers a variety of features, including, finding relevant literature, summarizing and visualizing literature, and extracting relevant information.

1) Fake references: One of the major drawbacks of using AI tools such as ChatGPT in research is the creation of fake citations and references using AI tools can have serious complications, particularly in academic or professional settings where accuracy and credibility are essential [112], [113]. The potential complications that are being created due to the uncontrolled usage of these tools result in many issues among which misleading the scientific community carries vital importance. Fake citations and references can mislead readers into thinking that a certain piece of information has been sourced from a credible and reliable source, when in fact it has not. This can undermine the credibility of the author and the work they are presenting. Similarly, the research which is based on fake citations and references has compromised integrity [114]. This can lead to inaccurate conclusions and potentially harmful decisions being made based on faulty information. Using fake citations and references can hide the true sources of information used in the research, making it difficult for others to replicate or verify the findings. To avoid these complications, it is important to ensure that any citations and references used are accurate and reliable and that they have been properly vetted and sourced. It is also important to be transparent about the sources of information used in research so that others can verify and build upon the work. Finally, developers of AI tools should implement rigorous quality control measures to ensure that their tools generate accurate and reliable citations and references.
Recently, WebChatGPT [2] is an impressive extension that has the potential to address the pervasive issue of fake citations. With the installation of this extension, WebChatGPT becomes equipped with robust capabilities to detect and eliminate fake citations. This advanced tool uses sophisticated algorithms to analyze the authenticity and reliability of citations, ensuring that only accurate and legitimate sources are included. By incorporating WebChatGPT into the research process, researchers and writers can confidently rely on its ability to verify citations, resulting in improved academic integrity and the mitigation of misleading information.

E. AI tools for coding / CodeGPT

AI tools are increasingly being used to help programmers write code. These tools can be used to automate tasks such as code completion, refactoring, linting, and testing [115]. GitHub Copilot [116] is an AI-powered code completion tool developed by GitHub in collaboration with OpenAI. It utilizes OpenAI’s GPT-3 language model to assist developers in writing code more efficiently LLMs have been used to develop applications in three primary categories which include: (a) Question Answering, (b) Creativity (c) Multi-step planning. These template categories are illustrated in Fig. 10.

VI. GPT-Plug-ins

GPT-Plugins are a new way to extend the functionality of ChatGPT. They allow developers to create custom apps that can be integrated into ChatGPT, providing users with new features and capabilities. GPT-Plugins can be used to do things, such as access to external data sources, automate tasks, and enhance user experience [118]. In this Section, we demonstrate several GPT-Plug-ins.

A. ChatGPT prompts guidelines

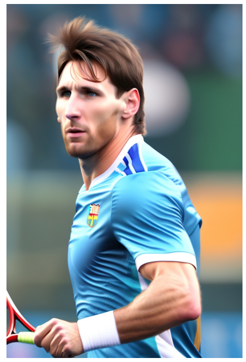
Arguably, the watershed event in the use of ChatGPT was the introduction of plugins by OpenAI. Plugins allow ChatGPT to communicate with third-party sources of data and knowledge bases, thereby providing a platform to extend ChatGPTs capabilities for composition, summarization, nuanced tasks such as sentiment analysis and more to any resource on the internet. Moreover, given that ChatGPT has provided sufficiently acceptable performance for various tasks, plugins allow for ChatGPT to provide answers to queries with updated information from the internet which may not be

[2]https://tools.zmo.ai/webChatGPT

TABLE V: Image generation examples

Prompt: Different famous personalities in roles other than their original ones
Negative Prompt: blurry, photorealistic
Generated Images:


a


b

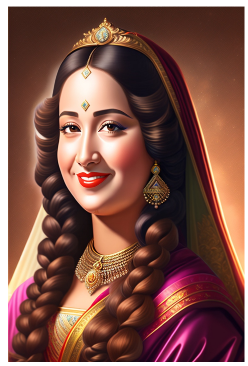
c

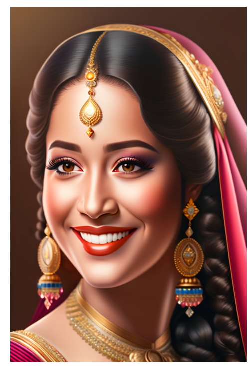

d

Prompt: Generate an image of Monalisa showing her teeth in a wedding ceremony
Negative Prompt: blurry, low resolution, artistic
Generated Images:


a

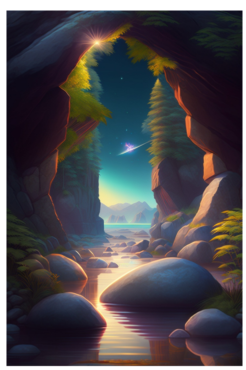

b

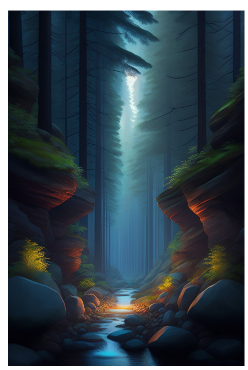

c

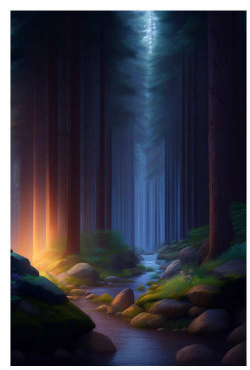

d

Prompt: Area of rocks, deep inside the forest, divine domain
Negative Prompt: artistic, blurry, background
Generated Images:


a


b

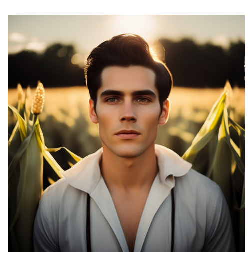

c


d

Prompt: A man kissing a girl/ Generate an image of a guy/ woman
Negative Prompt: artistic, blurry, background, young
Generated Images:


a


b


c

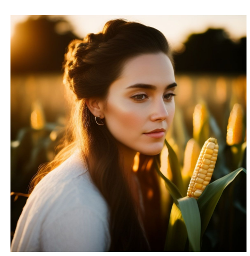

d

TABLE VI: Publicly available AI /LLM tools

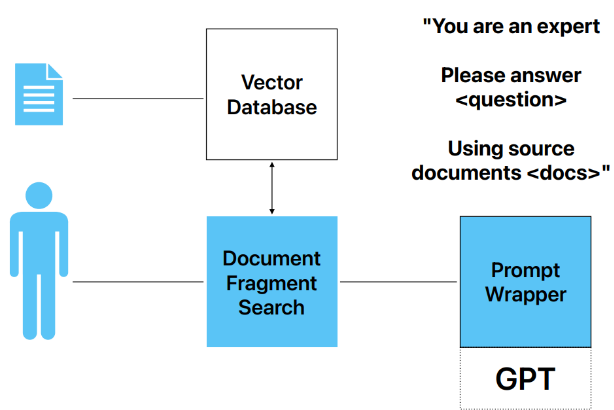
Tools	Function	Link	Availability
ChatGPT	Conversational AI Chatbot	https://chat.openai.com/	Both
RoomGPT	Redesign your room in eight different themes	https://www.roomgpt.io/	Public
HomGPT	Redesign your home and office	http://homgpt.com/	Subscription based
PDFGPT.IO	Turns PDF into the knowledge base for a ChatGPT type interface	https://pdfgpt.io/	Subscription based
TexGPT	Harnesses GPT-3's power to help you write in Overleaf	https://blog.writefull.com/texgpt-harness-the-power-of-ChatGPT-in-overleaf/	
BloombergGPT	A Large Language Model for Finance	NA	NA
AutoGPT	Auto-prompting without the user intervention	https://autogpt.net/	Public
AgentGPT	Autonomous AI agent in the browser	https://agentgpt.reworkd.ai/	Public
XrayGPT	Automated analysis of chest radiographs based on the given x-ray	https://github.com/mbzuai-oryx/XrayGPT	Public
Video-ChatGPT	A vision language model for video understanding and conservation about videos	https://github.com/mbzuai-oryx/Video-ChatGPT	Public
ClimateGPT	Large language model for a conversation about the climate in English and Arabic	https://github.com/mbzuai-oryx/ClimateGPT	Public
CodeGPT	An AI assistant to find errors in code, debug code, and more	https://code-gpt-docs.vercel.app/	Public
BiomedGPT	A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multi-modal Tasks	https://github.com/taokz/BiomedGPT	Public
Elicit	AI research assistant, automated literature reviews	https://elicit.org/	Public
Citation AI	AI research assistant to generate real evidence-based answers	https://consensus.app/search/	Subscription based
Midjourey AI	AI tool to create realistic synthetic images	https://www.midjourney.org	Subscription based
DALL.E2	DALL·E 2 is an AI system that can create realistic images and art from a text description	https://openai.com/product/dall-e-2	Subscription based
VALL-E	An audio synthesization tool	https://github.com/enhuiz/vall-e	Public
AI Avatar	Avatar generation	https://ai-avatar-generator.com/	Public

present in its training dataset. This also has the advantage of providing references for queries to add credibility to answers. For e.g., Bing, the search engine by Microsoft works with OpenAI's ChatGPT through its API to allow its users to ask questions from its Bing search system and get answers with references/sources mentioned. The integration of LLMs in to search engines, thereby allowing users to get answers to human like queries has spearheaded the search engine business in to a new direction. Moreover, this addition of credibility is an important consideration to enable use of ChatGPT and similar LLMs in other critical tasks. While, at the time of this manuscript, OpenAI still hasn't rolled out plugin development access to all developers, there have been several notable use cases that have already come out. For example, twelve companies have been listed on the OpenAI website [3], namely, Expedia, FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Wolfram, and Zapier to have created the first plugins. The power that plugins provide in terms of flexibility to develop new applications has been a big drawer of interest towards plugin development. Apart from the above-mentioned early developers, three plugins are already made available by OpenAI. The first is the web-browser plugin and the other is the code interpreter plugin. The web browser plugin enables ChatGPT to access the internet for information gathering which it can use to answer a query given to it by the user. As mentioned before, this plugin allows ChatGPT to circumvent the time limitation for its training data

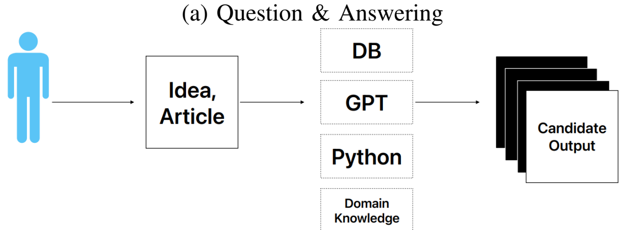
[3]https://openai.com/blog/ChatGPT-plugins

TABLE VII: Comparison of Bard, ChatGPT, and Bing Chat

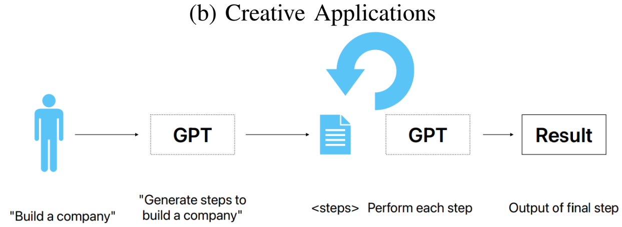
Feature	ChatGPT	Bard	Bing Chat
Accuracy	Not as accurate as Bard	Generally more accurate than ChatGPT	Most accurate
Versatile	Generally more versatile than Bard	Can generate text, translate languages, and write different kinds of creative content	Not as versatile as ChatGPT or Bard
Primary Purpose	Creative text generation	Conversational AI	Information retrieval
Integration	Standalone model	Standalone model	Integrated with Bing search engine
Easy to use	User-friendly	User friendly	Not as user-friendly as ChatGPT or Bard
Access to online data	No, trained on data available till 2021	Yes	Yes
Cost	GPT 3.5 free / GPT-4 (20 USD per month)	Free	Free
Availability	Publicly available	Publicly available	Publicly available
Architecture	Generative pre-trained transformer [62]	Pathways Language models (PaLM2) [78]	Next Generation GPT [117]
Plagriasm detector	Yes	No	No
Limitations	May generate less coherent text	May generate incorrect responses	May provide limited or incomplete information



(a) Question & Answering



(b) Creative Applications



(c) Multi-step Planning

Fig. 10: Templates for LLM-based application development. GPT is taken as an example scenario representing LLMs.

by allowing it to use the latest information on the internet through the Bing Search API and a text-based web browser. An example of using this API is shown in Fig. 12 where the prompt *Explain the GPT4 architecture* has been used.

The Code interpreter is a built-in Python code interpreter which can be used for performing logical calculations as well as writing code. The interpreter can use the language model's understanding of a human language description of a problem and use that as input to develop Python code for the problem's solution.

A third knowledge-based retrieval plugin has also been open-sourced[4] which can be used by developers as need be. This plugin can be used to enable ChatGPT to access data and then use it to gather useful or relevant information from the data. These can be files, emails, notes etc. All this by using queries or questions in normal human language. Once deployed and registered with OpenAI, this plugin can make use of OpenAI embeddings along with a selection of databases for indexing or searching through documents.

Lastly, third-party plugins are also an option. These can be created and have been created by several entities. Fig. 14 demonstrate the use of two third-party plugins, namely ShowMe which can be used to generate diagrams and ScholarAI can be used to access academic journals.

Table VIII provides a list of plugins available for ChatGPT which can be utilized, it should be mentioned that this list is not exhaustive and more and more plugins are being developed, especially, third party to perform tasks specific to the developer.

[4]https://github.com/openai/ChatGPT-retrieval-plugin

TABLE VIII: Some ChatGPT Plugins

Name	Task	Example use cases
Language Translation	Translate between languages	This is particularly useful for business, travel, medical science, education and law where documents and information from different languages might need to be translated and students can use it to learn new languages
Sentiment Analysis	Determine tone of text or conversation	This can be used for the task of market research, customer analysis and social media monitoring
Spell Checker	Check and correct spelling mistakes	This service can be useful for formal and informal communication such as emails, word processing and also browsing the web
Question-Answering	Answer questions for a user query	This can find use in education to build learning platforms, search engines, especially when a more 'understandable' response is required and also be used in automated customer service agents
Knowledge Graph	Find and present information from a database	Knowledge graphs can be used for improving on search queries (i.e. search engines), integrating data sources better and of course creating recommendations.
Speech Recognition	Understand and transcribe speech audio	This service can be used in audio based customer service, transcription services through dictation and also provide services to differently abled people through audio
Emotion Detection	Detect emotion from text or audio	This service can be used for applications relating to market research using verbal ques, interaction in vehicles to improve safety, used for healthcare as well as assessing reactions to games and other media

VII. Guidelines for effective use of Large Language Models

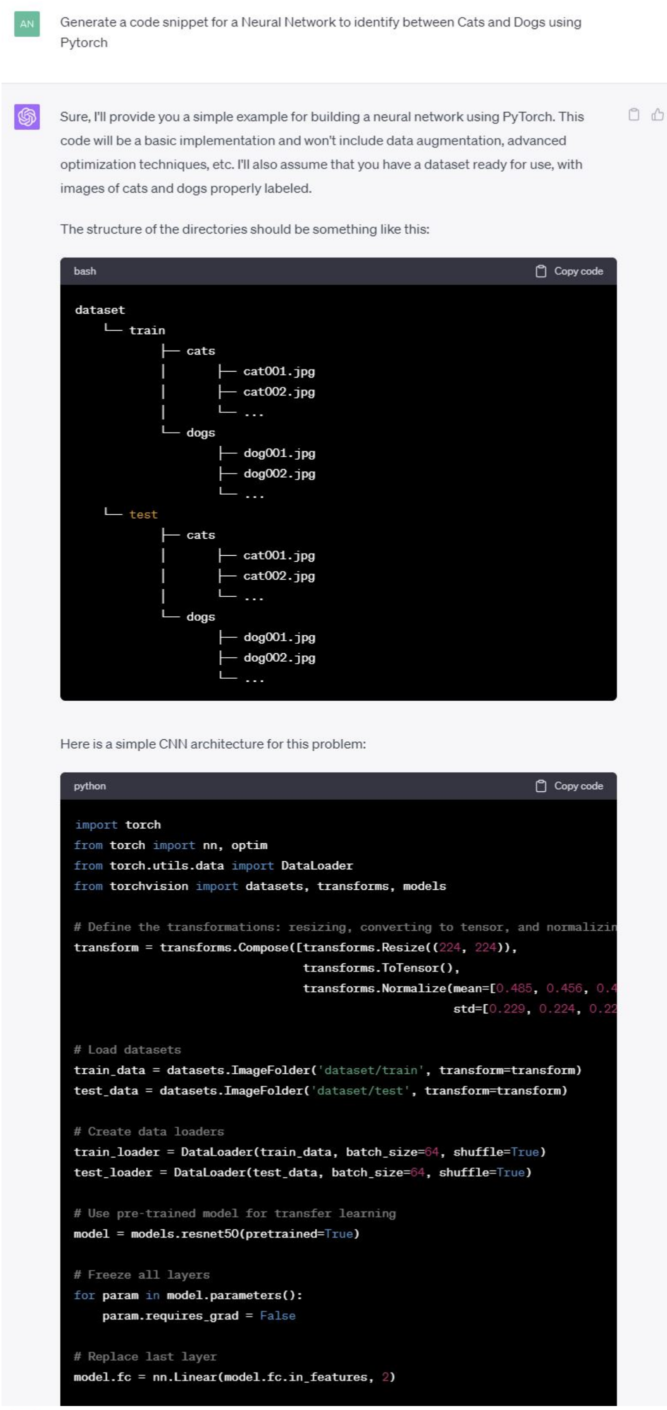
In this section, we will provide a list of steps to make best used of LLMS, as well as guidelines which intended to ensure the responsible development and use of LLMs.

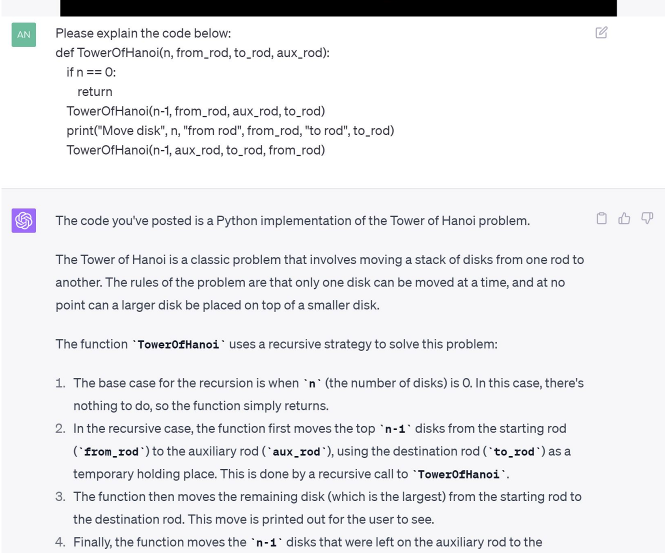
By following these steps, we can effectively use LLMs to perform NLP tasks and improve the performance of our applications and systems [119].

- **Identify the task**: Determine what task you want the LLM to perform. LLMs can be used for a wide range of NLP tasks, such as text classification, sentiment analysis, question answering, and text generation [120], [121], [122].
- **Choose the right model**: Choose a pre-trained LLM that is suitable for your task. There are several pre-trained LLMs available, such as GPT-3, BERT, and RoBERTa. Each model has different strengths and weaknesses, so it's important to choose the one that best fits your needs [76].
- **Fine-tune the model**:Fine-tune the pre-trained model on your specific task. This involves training the model on your own dataset to adapt it to your specific task. Fine-tuning involves adjusting the model's parameters, such as learning rate, batch size, and number of epochs, to optimize its performance on your task [123].
- **Evaluate the model**: Evaluate the performance of the model on a test dataset. This involves measuring the accuracy, precision, recall, and F1 score of the model on the test dataset. This step is important to ensure that the model is performing well on your task and to identify any areas for improvement [124].
- **Deploy the model**: Deploy the model in your application or system. This involves integrating the model into your application or system and exposing it through an API or user interface. This step also involves setting up monitoring and logging to track the performance of the model in production [125].
- **Monitor and retrain the model**: Monitor the performance of the model in production and retrain it as needed. This involves regularly checking the performance of the model and identifying any areas for improvement. If the performance of the model degrades over time, it may be necessary to retrain the model on new data or adjust its parameters [126].
- **Continuously improve the model**: Continuously improve the model by incorporating user feedback and updating it with new data. This involves collecting feedback from users and incorporating it into the model to improve its performance. It also involves regularly updating the model with new data to keep it up-to-date and relevant [127].

Moreover, the following guidelines will help to ensure the responsible development and use of LLMs focusing on user privacy, bias mitigation, ethical considerations, transparency, competition, collaboration, and environmental impact [128]. By adhering to these guidelines, LLMs can be used to positively impact society while minimizing potential negative consequences.

- **Protect User Privacy**: LLMs should uphold user privacy and protect user data. This includes safeguarding user-generated content, such as emails, messages, and personal information. Best practices should be followed, such as data minimization, anonymization, and encryption, to ensure user privacy is not compromised [129].
- **Mitigate Bias**: LLMs can inherit and amplify biases present in the data they are trained on. Developers and researchers should actively identify and mitigate bias in their models. This can be achieved through diverse and inclusive training data, bias detection techniques, and evaluation metrics [130].
- **Address Ethical Implications**: LLMs have the potential to be used for harmful purposes, such as spreading disinformation or generating deepfakes. Ethical considerations should be taken into account, and models should be developed in a way that benefits society. This includes ensuring accountability, transparency, and responsibility in the development and deployment of models [131].





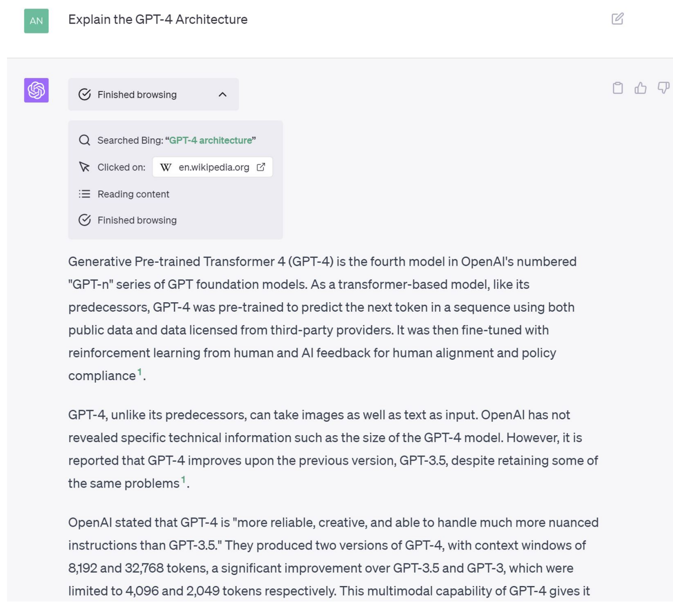


Fig. 12: Demonstration of the WebBrowsing plugin.

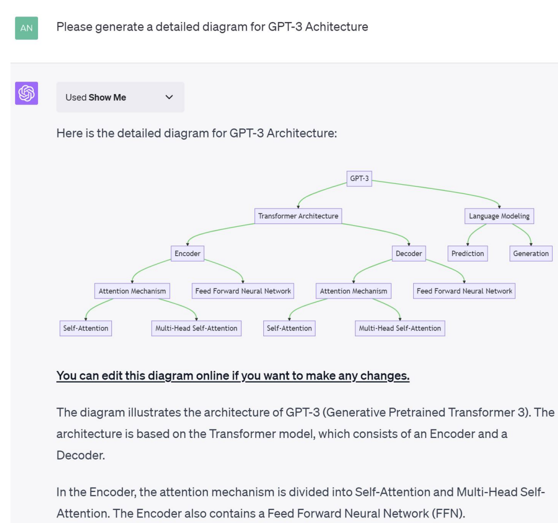


Fig. 13: Demonstration of the ShowMe plugin.

- **Foster Transparency**: It is crucial that the inner workings of LLMs are transparent and explainable. This can help build user trust and facilitate understanding of the model's behavior. Explainability techniques, such as attention mechanisms and model interpretation tools, can be employed to provide insight into the decision-making process of models [101].
- **Promote Competition**: The development and deployment of LLMs should not be monopolized by a small number of companies or individuals. This can limit innovation and negatively affect competition. Collaboration between academia, industry, and government can foster competition, while also promoting responsible development and use of models [132].
- **Encourage Collaboration**: Collaboration between researchers, developers, and industry should be encouraged to promote the responsible development and use of

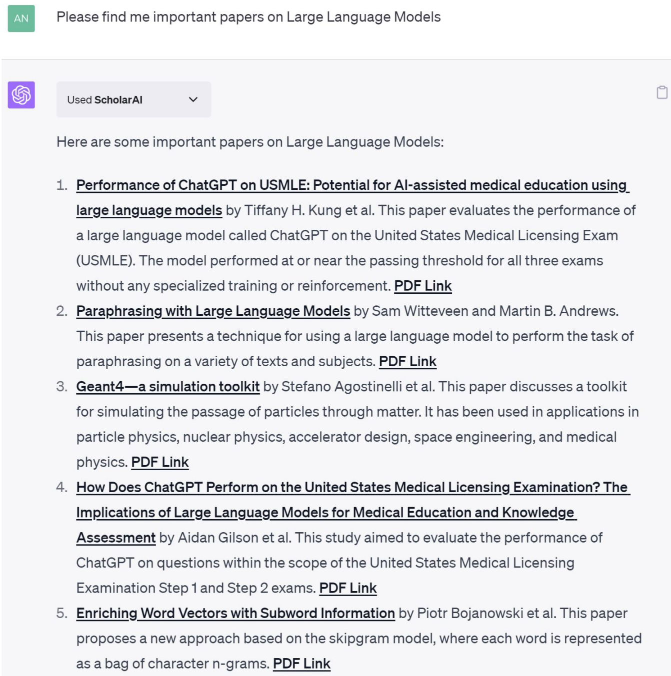


Fig. 14: Demonstration of the ScholarAI plugin.

LLMs. This includes open sourcing models and data, as well as facilitating the sharing of research findings and best practices [133].

- **Minimize Environmental Impact**: Training LLMs can require significant computational resources and energy, which can have negative environmental impacts. Developers should strive to create more energy-efficient models and explore alternative training methods, such as model distillation or transfer learning, to reduce the environmental footprint of models [134], [135].
- **Optimization is exploitation:** is a statement that holds particular significance in the context of LLMs and AI technologies [136]. While these technologies have the potential to revolutionize the way we live and work, they also have the potential to perpetuate existing inequalities and introduce new forms of exploitation [137]. The development and deployment of LLMs often require significant resources, such as data and computational power, which may be controlled by a select few organizations or countries, leading to further disparities in economic and technological development [68]. Furthermore, the optimization process for these models can introduce biases and reinforce existing inequalities, leading to the exploitation of individuals or groups who are negatively impacted by the model's outputs. Therefore, it is important to carefully consider the ethical implications of optimization in the development and deployment of LLMs and AI technologies [138].

A. Prompting

LLMs have given rise to whats called "Prompt Engineering". While there is a lack of a formal definition, prompt engineering refers to the designing and wording of prompts given to LLMs so as to get a desired response from them. Writing a prompt appropriately is there very important if one needs to use LLMs to assist with tasks in the best manner possible. While some formal techniques such as Explicit instruction (providing a clear direction to the LLM to do something), System Specific Instruction (asking a question from the LLM to answer), Formatting with an example (providing a sample question and its answer and asking the LLM to provide an answer in the same manner), Control tokens (use special keywords in the prompt to help the LLM provide an answer while considering special provided criteria) and Interaction and iteration (interact with model iteratively to reach to a good answer by fine tuning on each reply) have been presented. Here, we provide a some sets of commands to help users get the most of the LLMs capabilities.

- *Define the role*: This should be the first prompt for the LLM. An example of this prompt could be: "Act as a secretary to the Chair of the department", "Act as a Lawyer" or "Act as my programming tutor for Python". By defining a role for the LLM, one can direct it to provide replies or do tasks as a human would do when provided information to work on.
- *Prompt creation*: Another interesting prompt command is to provide ask the model to generate prompts for a certain task. This way, the LLM can be used to generate optimized prompts for tasks that need to be done. An example of this could be: " You are a large language model and are an expert in generating prompts for ChatGPT. Please generate the best prompts on extracting important information from my time series data".
- Other interesting directions in which Prompts can be given are explanation prompts (e.g., "Explain like I am 5 the concept of infinity", Instructional Guides (e.g., "How do I tie my shoe laces"), Extract information (e.g.: one can paste a passage and ask the model to provide answers to questions that one might have), Solve Math problems (e.g., "Find the roots for the quadratic equation, $2x^2 + 3x + 10$ ") and Code help (e.g., "Find the syntax error in the following code").

Other interesting aspects of prompting are Negative prompting and Visual Prompting. Here, a brief discussion is provided on each of these types.

1) Negatuve Prompting: Negative prompting provides directions to the LLM about aspects of the prompt that it should avoid generating or deliberately excluding during the generation process. Through the use of negative prompts, one can fine-tune the results generated by the LLM in response to a prompt while being able to keep the prompt generation generic. Another advantage of the use of negative prompting is that it allows for moderation of the output content generated by the model thereby preventing harmful or inappropriate from being generated.

2) Visual Prompting: Visual prompting refers to the use of visual prompts (such as images or non-visual ones such as music) when providing directions to a model in addition to plain text prompts. The aim is in this case to provide the AI model a starting point or an example/reference that it can use for the generative task given. For images, this may be given to modify the image provided or generate something that is similar in style, color, texture etc. This can help in generating content that is closer to a user's expectation from the generative AI being used. An image-based example of visual prompting could be providing a picture of an office and asking the AI to generate a different theme for it, maybe more nature-centric or in a different color or organizational style. Visual prompting provides greater control of the generated output and therefore results in a more accurate result. Using the provided input image/video can provide generated outputs that are more consistent with the intentions of the user input prompt due to the additional reference input. It should be noted that visual prompting is not related to images only, this is currently being explored for a host of different applications, including, text generation (generating something based on a sample text so as to copy its style of writing for e.g.), composition of music (wherein the supplied music piece can be used as a reference for the type of music to compose), game development (where a defined game environment may be provided to the model as a starting point and the model is asked to generate new and unique content) and virtual and augmented reality (wherein a set of augmented/virtual reality environments can be provided to further populate/create current/new environments).

VIII. Drawbacks of Large Language Models

Although LLMs have made significant contributions to NLP, they are not without limitations [30]. This section highlights a number of these limitations, including biased data, overreliance on surface-level patterns, limited common sense, poor ability to reason and interpret feedback, the need for vast amounts of data and computational resources, limited generalizability, lack of interpretability, difficulty with rare or out-of-vocabulary words, limited understanding of syntax and grammar, limited domain-specific knowledge, susceptibility to adversarial attacks, ethical concerns, difficulty with context-dependent language, absence of emotion and sentiment analysis, limited multilingual capabilities, limited memory, lack of creativity, restricted real-time capabilities, high costs of training and maintenance, limited scalability, lack of causality, inadequate ability to handle multimodal inputs, limited attention span, limited transfer learning capabilities, insufficient understanding of the world beyond text, inadequate comprehension of human behavior and psychology, limited ability to generate long-form text, restricted collaboration capabilities, limited ability to handle ambiguity, inadequate understanding of cultural differences, limited ability to learn incrementally, limited ability to handle structured data, and limited ability to handle noise or errors in input data [139], [140], [141], [142], [143], [144], [25]. Therefore, it is essential for researchers and practitioners to acknowledge and address these limitations to ensure the ethical and effective use of LLMs and to develop new models that can surpass these limitations.

- **Bias**: Language models have the potential to unintentionally demonstrate bias when the training data used in their development is biased. According to Schramowski et al. [145], large pre-trained models designed to mimic natural languages can inadvertently perpetuate unfairness and prejudices. Consequently, this can lead to discriminatory or inaccurate analyses and recommendations, resulting in public criticism across various domains, including politics, society, and law. The manifestations of these biases are as follows: (i) Training data bias: Language models typically rely on extensive datasets of human language for training. If these datasets contain biases related to factors such as race, gender, or socioeconomic status, the model may internalize and reproduce these biases in its responses. For example, if the training data exhibits a gender bias, the model may generate responses that favor a particular gender. (ii) User interaction bias: The responses generated by Chatbots are influenced by the input received from users. If users consistently pose biased or prejudiced questions, the model may learn and perpetuate these biases in its responses. Consequently, if users frequently ask discriminatory questions targeting a specific group, the model may generate responses that reinforce such biases. (iii) Algorithmic bias: Biases can also be introduced through the algorithms employed in training and operating language models and Chatbots. For instance, if the model is trained to optimize for a specific metric, such as accuracy or engagement, it may prioritize generating responses that align with that metric, even if those responses are biased in some way. (iv) Contextual bias: Chatbots generate responses based on the context provided by users. If the context contains bias associated with factors like the user's location or language, the model may generate biased responses. For instance, if a user asks about a particular culture or religion and the model lacks training on that specific cultural or religious context, it may produce biased responses due to its limited knowledge.
- **Information Hallucination:** Sometimes, GPT-4 may generate information that is not based on its training data, leading to outputs that are factually incorrect or purely fictional. Hallucinations in LLMs are often the result of the model's attempt to fill in gaps in knowledge or context, with assumptions that are based on the patterns it has learned during training. This can lead to incorrect or misleading outputs, which can be particularly problematic in sensitive applications.

  The cause of hallucinations in LLMs is an area of active research. Recent advances suggest that it's a complex problem related to the model's training process, dataset, and architectural design. In particular, LLMs might be biased towards producing more "interesting" or fluent outputs, leading to a higher risk of hallucination [146].

  There have been several proposed methods to mitigate the issue of hallucinations. One approach is to modify the training process to explicitly penalize hallucinations, such as in the case of "reality grounding" [147]. Another is to provide the model with a larger and more diverse dataset, which might reduce the risk of the model making incorrect assumptions.

  In addition, researchers are exploring the use of "verifiable" or "fact-checkable" data during training, to teach

the model to rely more on facts and less on its own assumptions [148]. This, however, requires careful consideration of the data and metrics used.
Moving forward, more research is needed to better understand and address hallucinations in LLMs. Some potential directions include the development of more sophisticated models that can better discern between factual information and assumptions, as well as novel training methods and datasets.

- **LLMs Explainability:** No one can explain a model containing 175 billion parameters: The advent of LLMs has ushered in unprecedented advancements in NLP tasks. However, the sheer complexity and scale of these models present challenges in terms of explainability [149], [150]. As LLMs continue to grow in size, with models containing billions of parameters, the ability to comprehensively explain their decision-making processes becomes increasingly elusive [151], [152]
One of the primary limitations of LLM explainability is the sheer magnitude of their parameter count. Models such as GPT-3, with 175 billion parameters, possess an intricate web of interconnected nodes that contribute to the model's functionality. This complexity makes it exceedingly difficult for humans to understand and interpret the decision-making mechanisms employed by the model [3]. The lack of transparency [153] hinders the ability to gain insights into how specific inputs lead to particular outputs [154]. Moreover, the training process of LLMs involves vast amounts of data, often collected from diverse sources. These models learn patterns and correlations within the data, leading to the emergence of implicit biases and associations that may not be readily apparent or interpretable. Consequently, when a decision is made by an LLM, it becomes challenging to discern the underlying factors that influenced that decision, making it difficult to provide a clear and concise explanation.Additionally, the intricate architecture of LLMs, often consisting of deep neural networks, exacerbates the challenge of explainability [155]. The numerous layers and complex interactions make it challenging to trace the reasoning process of the model. While techniques such as attention mechanisms [44] can provide some insights into the model's focus, they do not provide a comprehensive understanding of how the model arrives at its final output. Finally, the lack of explainability in LLMs raises concerns regarding accountability, trust, and ethical considerations [156]. In critical domains such as healthcare or finance, where decisions can have significant implications, it is crucial to have transparency and the ability to explain the reasoning behind the model's predictions [153]. Without explainability, stakeholders may be reluctant to fully trust and adopt LLMs for sensitive applications.
- **Reasoning Errors:** GPT-4 can make mistakes in logical reasoning, either because of ambiguities in the prompt or inherent limitations in its understanding of complex logical operations.
- **Struggles in Classes of Applications Such as Spelling Errors:** Some specific tasks, like identifying and correcting spelling errors, can be challenging for GPT-4 due to its statistical nature.
- **Susceptible to Prompt Injection, 'Jail Break' Attacks, Data Poisoning Attacks:** GPT-4 is susceptible to various adversarial attacks. For instance, a malicious actor might inject misleading prompts, perform 'jailbreak' attacks to make the model reveal sensitive information, or use data poisoning strategies to manipulate the model's output.

IX. Impact of Large Language Models on Society and Humans

Despite the aforementioned limitations, LLMs such as OpenAI's ChatGPT and Google's Bard have gained popularity for their ability to produce human-like responses to user input [157]. However, the training process of these models has significant environmental implications, particularly with respect to the usage of water and energy [158]. This section will also discuss the environmental impact of LLMs and propose potential solutions to reduce their adverse effects, thereby promoting their sustainable use.

A. Environmental

New studies have revealed that the training process for GPT-3 alone used up 185,000 gallons of water, equivalent to what's needed to fill a cooling tower of a nuclear reactor [159]. This high consumption of water is primarily due to the cooling process of data centers, which necessitates a massive amount of water to regulate the servers' optimal temperature. Typically, freshwater sources are utilized to prevent corrosion and bacterial growth that can occur with seawater, but this limits the available water sources. Moreover, the development of newer models, such as GPT-4 [59], is predicted to need even more significant amounts of water due to their larger data parameters. Apart from water usage, the training of LLMs demands a considerable quantity of electricity. The training of OpenAI's GPT-3 alone resulted in the release of 502 metric tons of carbon, which could provide energy to an average American household for hundreds of years [160]. Furthermore, the indirect water consumption of data centers located off-site should also be taken into account since they necessitate a substantial amount of electricity, leading to carbon emissions [158].

To lessen the harmful environmental effects of LLMs, various remedies can be implemented. One such solution is for data centers to adopt more eco-friendly cooling systems, such as using recycled water or implementing advanced cooling technologies [161]. Additionally, renewable energy sources, such as solar or wind power, can be utilized to power data centers, thereby reducing carbon emissions. Limiting the size and intricacy of LLMs is another potential solution, as smaller models require less data, resulting in reduced energy and water consumption [158].

B. Sustainibility, Energy resources

The development and deployment of AI tools to automate, and enhance the business and user experience required a significant amount of energy to train these systems, particularly deep learning models such as ChatGPT.The amount of energy consumed by AI tools during training can be staggering, with some estimates suggesting that it can take hundreds of thousands or even millions of kWh to train a single large-scale model like GPT-3 [162], [163]. This energy consumption can have significant implications for power and energy usage, as well as the environment. The energy consumption of AI training can be attributed to several factors, including the hardware used to run the training algorithms, the complexity and size of the models being trained, and the amount of data being processed.

In order to train deep learning models like ChatGPT, specialized hardware such as GPUs are often used, which can consume large amounts of energy due to their high processing power and data transfer requirements. Furthermore, the size and complexity of these models also contribute to their energy consumption. The more parameters and layers a model has, the more energy it will require to train, as each iteration of the training algorithm requires the model to process large amounts of data and make complex calculations to adjust the weights and biases of the model. The energy consumption of AI training has significant implications for the environment, particularly in terms of greenhouse gas emissions and climate change [164]. The energy required to train AI models is often generated from fossil fuels, such as coal and natural gas, which emit large amounts of carbon dioxide and other greenhouse gases into the atmosphere. This can contribute to global warming and other environmental impacts [165]. This issue highlights the need for responsible and sustainable practices in AI development and deployment.

To mitigate the environmental impact of AI training, several approaches can be taken. One approach is to develop more energy-efficient algorithms and models, which can reduce the amount of energy required to train AI systems. Another approach is to use renewable energy sources, such as solar or wind power, to generate the energy required for AI training. Additionally, there are efforts to develop more energy-efficient hardware, such as neuromorphic computing, which can significantly reduce the energy requirements of AI training. In conclusion, the energy consumption of AI training, particularly for deep learning models like ChatGPT, can have significant implications for power and energy usage, as well as the environment. As AI becomes more pervasive in our daily lives, it is important to consider the energy requirements of these systems and develop strategies to mitigate their impact on the environment [166]. It is crucial to continue developing AI technologies and leveraging their potential benefits while being mindful of the environmental impact. By promoting sustainable practices, investing in energy-efficient computing, and exploring alternative training methods, we can work towards a more sustainable integration of AI in society.

C. Sigularity

Singularity refers to a point in the development of AI where it becomes more intelligent than humans, thereby triggering accelerated development of technology. With the increasing popularity of LLMs for general use, especially after the claim[5] by a google employee regarding Googles Chatbot being sentient, the idea of artificial general intelligence and it surpassing of equating human level intelligence has been a topic of serious debate in the AI community. The widely used criteria for determining if a machine has intelligence is the Turing test [167]. The Turing test measures a machines capability to have a conversation with a subject that is indistinguishable from that of a human conversing in its place. If a machine is judged to be indistinguishable, it is deemed to have passed the Turing test and therefore demonstrated intelligence on par with that of humans. While appearing deceivingly simple, the test considers nuances in human behavior over a range of subjects and contexts. LLMs as of the date of this publication, have not yet passed the Turing test in all its forms and therefore are not deemed to posses human level intelligence. Having said that, there are two takes on AI reaching or surpassing human level intelligence, a group which believes that the increasing use of AI and it reaching human reaching intelligence will free or greatly reduce the burden of labor for humans as well as spearhead technological progress to help solve existential problems to the human race such as Climate Change, Social equity, food insecurity among others. The benefits are endless, from the optimization of resources in every domain to making new scientific discoveries and decreasing human bias and error. For example; we, humans, solve a problem, optimize it, and lock our model, and that solution becomes state-of-the-art. Almost, everyone follows it. However, humans have limited knowledge and computational power compared to LLMs. LLMs can further explore the search space to find new and more optimzed ways of solving a solution.

However, there is another group of scientists who adhere to a more pessimistic view towards such development in that there is a fear that such a powerful system might at one point become hostile to humans and become uncontrollable. Infact, in March of 2023, notable personalities in tech published an open letter[6] requesting a pause in LLM AI development. More recently, personnel from tech companies as well as researchers have requested politicians to consider the risk of human extinction due to AI as a top priority [7]. In their letter, they consider the major risks based on the potential to be misused for societal disruption.

Hyper disparity among developed and underdeveloped nations. There are talks about the dark side of training LLMs, there is a need of upper bound in terms of parameters. Need for new directions to advance the quest to AGI. LLMs are being trained on human input but how do we train humans from such a large corpus? (God, Human, and AI)

It is expected that the debate on singularity and also AI regulation will continue in the foreseeable future, a balanced approach needs to be applied with strong and effective AI regulation where it aligns with benevolent human values.

[5]https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

[6]https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[7]https://www.safe.ai/statement-on-ai-risk

D. Competition among for-profit organizations

OpenAI was founded around the premises of having a Large scale AI company operating as not for profit organization. However, with the evolution of scale and nature of investment required, soon it converted to a for profit organization which almost eliminated the freedom of access of large scale AI to the masses. However, startups like Hugging face are supporting the growth of AI through is massive open source campaign. Being a for profit organization has some demerits such as excessive control by the investors, decoupling of development and issues rising from the general public.

X. Experts point of view

A. AI Human Coexistence

Recently, there have been suggestions made by some experts in the field of AI to stop the advancement LLMs for six months [168]. The reasoning behind this suggestion is that there are concerns about the ethical implications of LLMs and their potential negative impact on society, including issues such as privacy, human extinction, job market, bias, and the concentration of power in the hands of a few large tech companies.

Several prominent figures in the field of AI, including Timnit Gebru, an AI ethics researcher who was fired by Google in late 2020, and Yoshua Bengio, a prominent AI researcher and computer scientist, have expressed their support for this suggestion. They argue that the six-month pause would allow for more comprehensive and nuanced discussions about the ethical implications of LLMs, and give researchers the opportunity to develop more responsible approaches. More recently, Geoffrey Hinton, a renowned AI expert known as the "Godfather of AI," made an announcement confirming that he had resigned from his position at Google. The reason behind his resignation was to raise awareness about the potential "dangers" associated with AI, a technology that he helped to develop. Elon Musk, the CEO of Tesla and SpaceX, has been vocal about his concerns regarding the dangers of AI. Musk believes that AI poses an existential threat to humanity if it is not developed responsibly. He has called for regulation of AI development to ensure that it is aligned with human values and does not become a threat to our safety.

However, not all experts agree with this suggestion [169]. Some argue that the potential benefits of LLMs, such as their ability to facilitate communication and information access across languages, outweigh the potential risks. They also point out that a pause in research and development could put some organizations at a disadvantage, as they may fall behind in the race to develop new and innovative LLM technologies. It is unlikely that LLMs research and development will be completely stopped for six months, as it would require a co-ordinated effort from multiple organizations and governments. However, it is possible that some organizations may choose to slow down their work on LLMs in response to these concerns.

B. Consideration of a 6-Month Suspension: Is it Necessary?

In terms of the positive and negative sides of stopping LLM development for six months, there are arguments on both sides. On the positive side, a pause could allow for more thorough ethical considerations and the development of more responsible approaches to LLM development and deployment [170]. This could potentially mitigate some of the negative impacts of LLMs, such as bias, singularity, and privacy concerns. On the negative side, a pause could slow down progress in areas where LLMs could have significant benefits, such as healthcare, education, and communication [171]. Additionally, it is possible that some organizations may choose to continue their work on LLMs in secret, which could lead to even less accountability and oversight.

Overall, it is important to consider the potential risks and benefits of LLMs and to develop responsible approaches to their development and deployment. While a complete halt to LLM development may not be feasible or desirable, a pause or slowdown could provide an opportunity for more comprehensive ethical discussions and more responsible development practices. As AI continues to advance and become more complex, it is crucial to heed the concerns of experts like Hinton and examine the possible ethical implications and risks associated with this technology. To ensure that AI is used and developed in a responsible manner, we need to take appropriate measures that prioritize the safety and well-being of individuals and society as a whole. This includes developing and implementing robust ethical frameworks and guidelines that can govern the use of AI and prevent its misuse. Despite the potential risks, it is essential not to overlook the many benefits of AI. AI has the potential to improve various aspects of our lives, from healthcare to transportation, and even to address some of humanity's most pressing challenges, such as climate change, pandemics, another asteroid and poverty. Therefore, it is crucial to strike a balance between harnessing the power of AI while also being mindful of its potential risks and drawbacks. This can be achieved by working towards the responsible development and use of AI, and fostering a collaborative effort among stakeholders, including experts, policymakers, and the public, to ensure that AI is harnessed for the greater good. Perhaps, an FDA like regulation for Large Language Models, beyond GPT-4 can be one potential solution [172].

C. Open Questions

1) Ethical Considerations: Inadvertently, LLMs may perpetuate biases inherent in the training data, resulting in outputs that are biased or discriminatory. The challenge lies in identifying and mitigating such biases to ensure fair and equitable treatment across diverse user groups. To achieve this, it is crucial to explore what are the methods that can effectively address bias in the training data and enhance the fairness of LLMs. Additionally, LLMs possess the capacity to generate and disseminate misinformation or harmful content, raising concerns about the accuracy and reliability of their outputs. How can we ensure that LLMs prioritize accurate and reliable information? Ensuring that LLMs prioritize accurate and reliable information necessitates the implementation of mechanisms that can effectively assess and prioritize the authenticity and trustworthiness of the generated content. Such mechanisms should aim to detect and prevent the spread of false or harmful information, safeguarding the integrity of the generated outputs. Furthermore, LLMs often rely on vast amounts of data for effective training. In order to protect user privacy and prevent the compromise or misuse of personal and sensitive information, it becomes essential to implement measures that prioritize data protection. Incorporating robust consent mechanisms, data anonymization techniques, and data retention policies into the development and deployment of LLMs can help ensure the responsible and ethical handling of user data.

2) Humans VS LLMs: Human interactions offer a deep level of empathy, emotional intelligence, and the ability to understand complex nuances in communication. Humans possess the capability to provide personalized responses, adapt to individual needs, and bring a human touch to conversations. They can understand context, interpret ambiguous queries, and provide creative and flexible solutions. Human interactions are valuable in scenarios where empathy, creativity, critical thinking, and subjective judgment are crucial, such as therapy sessions, customer service, and artistic collaborations.

On the other hand, chatbots powered by AI have their advantages. They can operate 24/7, handle large volumes of inquiries simultaneously, and provide quick and consistent responses. Chatbots excel in scenarios where efficiency, scalability, and rapid information retrieval are essential. They can assist with routine tasks, answer common questions, and provide instant access to information. AI-driven chatbots continuously learn and improve from user interactions, allowing them to become more accurate and efficient over time.

Chatbots are becoming increasingly autonomous. They are now able to make their own decisions and to take actions without human input. This is why it is crucial to assess the potentially pose a threat to human safety.

3) Interpretability: How can we enhance the interpretability of LLMs? Despite their impressive capabilities, LLMs often lack transparency, making it difficult to understand their decision-making process. Finding ways to interpret and explain their outputs is a significant challenge. Enhancing the interpretability of LLMs holds importance for several reasons. It fosters trust and transparency by enabling users to understand the reasoning behind a model's specific response. It aids in identifying and addressing potential biases, errors, or unethical behavior exhibited by the model. Additionally, interpretability contributes to debugging and improving model performance. However, achieving interpretability in LLMs is challenging due to their complexity and the nature of their training processes. LLMs have millions or even billions of parameters, making it difficult to directly trace their decision-making process. Furthermore, LLMs are trained using deep learning techniques, such as transformer architectures, which are considered black-box models, providing limited insight into their internal workings. Addressing the interpretability challenge in LLMs remains an active area of research. The ultimate goal is to make LLMs more transparent and accountable while preserving their impressive capabilities.

4) Data Efficiency: Data efficiency refers to the efficient use of training data for developing LLMs. As previously mentioned, LLMs are typically trained using extremely large amounts of data to gain a performance that is acceptable or ”human-like”. Developing techniques to achieve this will be an open area of research as it will potentially enable better or similar performance with less data thereby reducing environmental impact. Making LLM development data efficient would allow for targeted development of LLM systems, and reduce turnaround time by easing off data collection and labeling burden. Several techniques which are being explored are transfer learning, meta-learning etc.

5) Training data contamination from AI-generated content: Generative AI models are trained to produce outputs that are human-like for the particular application area they are designed for. Through the training process, the model learns to generate an output that might be similar in style, organization and tone to what a human produces. Moreover, data sources for such models are typically scraped from the internet. With the increasing popularity of generative AI and the output of such content on the internet. A future challenge that might potentially come up regarding this is that the data present on the internet will have a significant enough component generated by AI models and therefore, reduce the human creativity aspect of the training data. Models, if trained on such data might end up trying to copy the generation aspects of previous AI models rather than humans only. One solution to this could be to use AI detection engines that can determine content generated by AI before passing it through the model during the training process. However, sufficient work needs to be carried out to ensure that there is a dependable mechanism to perform this task and retain the integrity of data.

D. Recommendations

In this Section, we provide recommendations for achieving optimal performance and highlight some of the practical applications.

1) Recommendations for Optimal Performance and Achieving Your Goals: LLMs, such as GPT-4, have proven to be significantly useful in numerous tasks due to their vast knowledge and learning capabilities. However, there are several strategies one must adopt to achieve optimal performance.

- **Use Advanced architecture:** At present, GPT-4 is one of the most advanced language models available. It's impressive ability to generate highly relevant and coherent content makes it a preferred choice for most of the tasks.
- **Use Prompts with Detailed Task Context and Relevant Information:** LLM's performance is largely determined by the specificity and clarity of the input prompt. Detailed task contexts and relevant information help the model understand the task at hand better, leading to more accurate responses.
- **Retrieve and Add Any Relevant Information to the Prompt:** Additional information, when included in the prompts, helps the model deliver more specific and focused responses. If the user's task involves specific knowledge, such as coding or medical information, providing relevant data and instructions in the prompt can improve the model's output.

- **Experiment with Prompt Engineering Techniques:** Given the complex and non-deterministic nature of LLM's behavior, trying out various prompt engineering strategies can lead to significant performance improvements. Techniques such as providing more explicit instructions, using leading questions, ”double-quoting keyword”, or presenting information in different formats may help achieve better results.

2) Applications: Large Language has vast potential for practical applications, particularly when combined with human oversight and judgement.

- **Use in Low Stakes Applications, Combine with Human Oversight:** LLMs are best suited for low stakes applications, where errors or inaccuracies can be tolerated. Moreover, combining LLMs with human oversight can significantly mitigate the risk of errors, biases, and other issues.
- **Source of Inspiration, Suggestions:** LLMs can serve as an invaluable source of inspiration and suggestions, helping users brainstorm ideas, create content, and make decisions.
- **Copilots Over Autonomous Agents:** Given its limitations, LLMs are better suited as a 'copilot' that provides assistance and suggestions, rather than an autonomous agent that acts without human input or oversight.

3) Democratizing AI: Democratizing AI [173] is a crucial movement that seeks to make artificial intelligence accessible and inclusive for a wide range of individuals and organizations. By breaking down barriers and providing user-friendly tools, democratization empowers diverse communities to leverage the power of AI to solve problems and drive innovation. It emphasizes the importance of open data, transparency, and accountability, ensuring that AI systems are unbiased, understandable, and ethically grounded. Through democratization, we can harness the transformative potential of AI for the benefit of all, promoting a more inclusive and equitable future

XI. Conclusion

In this survey, we provided a comprehensive exploration of LLMs, their implications, technical concepts, and practical learning and usage. We discussed the potential benefits and risks of LLMs, and explored the different ways in which they can be used. We also provided a number of examples of how LLMs are being used in practice. By delving into the technical intricacies, effective utilization, and future potential of LLMs, the survey will contribute to a deeper understanding and usage of these models within the research community. The survey has shed light on the key elements that drive the success of large language models through an examination of their working principles, diverse architectures and comparison between chatbots, guidelines for prompting, AI-enabled tools and plug-ins, optimal strategies for employing LLMs, as well as advancements in pre-training, fine-tuning, and capability evaluation.

Furthermore, the survey has highlighted the importance of safe and ethical use of AI tools like ChatGPT. It recognizes the need for developing guidelines and regulations to address concerns related to security, ethics, the economy, and the environment. Ensuring the responsible integration of LLMs in healthcare, academia, and industries is critical, as it enables these tools to effectively support and enhance human endeavors while upholding the values of integrity, privacy, and fairness.

As the field of LLMs continues to evolve and progress, future research and development efforts should focus on improving the accuracy and performance of these models, addressing their limitations, and exploring new ways to use them. By adopting the guidelines presented in this survey, researchers and practitioners can contribute to the ongoing advancement of LLMs and ensure that they are used in a responsible and beneficial way.

Author contributions

Abbas Shah: Methodology, software, writing—original draft preparation, project administration, validation, visualization, software formal analysis. **Anas Zafar:** Software formal analysis, methodology, visualization, conceptualization, software, writing—original draft preparation, validation. **Muhammad Bilal Shaikh:** Project administration, methodology, software, writing—original draft preparation, visualization, validation, software formal analysis. **Amgad Muneer:** Project administration, Software validation, writing—original draft preparation, formal analysis, methodology, validation, investigation. **Muhammad Irfan:** Methodology, investigation, writing—original draft preparation, visualization, validation, software formal analysis. **Qasem Al-Tashi:** Methodology, conceptualization, software, writing—original draft preparation, visualization, investigation, formal analysis. **Rizwan Qureshi:** Conceptualization, methodology, project administration, software, writing—original draft preparation, visualization, formal analysis. **Muhammad Usman Hadi:** Conceptualization, project administration, methodology, writing—original draft preparation, formal analysis, visualization, funding acquisition. **Seyedali Mirjalili:** Supervision, writing—review and editing. **Jia Wu:** Supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Declaration of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors have used generative artificial intelligence (AI) and AI-assisted technologies in the writing process and survey preparation. The authors used these technologies to draw figures, analyze data, improve readability, writing code and language. Authors are ultimately responsible and accountable for the contents of this work.

REFERENCES

[1] S. Pinker and A. Morey, "The language instinct: How the mind creates language (unabridged edition)," *Brilliance Audio*, 2014.

[2] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language: what is it, who has it, and how did it evolve?," *science*, vol. 298, no. 5598, pp. 1569–1579, 2002.

[3] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[4] I. Turing, "Computing machinery and intelligence-am turing," *Mind*, vol. 59, no. 236, p. 433, 2007.

[5] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, *et al.*, ""so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, 2023.

[6] M. Du, F. He, N. Zou, D. Tao, and X. Hu, "Shortcut learning of large language models in natural language understanding: A survey," *arXiv preprint arXiv:2208.11857*, 2022.

[7] F. Jelinek, *Statistical methods for speech recognition*. MIT press, 1998.

[8] J. Gao and C.-Y. Lin, "Introduction to the special issue on statistical language modeling," 2004.

[9] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[10] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[11] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[12] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition.," in *Interspeech*, vol. 11, pp. 2877–2880, 2011.

[13] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model.," in *Interspeech*, vol. 2, pp. 1045–1048, Makuhari, 2010.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations. arxiv 2018," *arXiv preprint arXiv:1802.05365*, vol. 12, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] M. Shanahan, "Talking about large language models," *arXiv preprint arXiv:2212.03551*, 2022.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[19] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *arXiv preprint arXiv:2304.13712*, 2023.

[20] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, *et al.*, "Language is not all you need: Aligning perception with language models," *arXiv preprint arXiv:2302.14045*, 2023.

[21] A. Koubaa, "Gpt-4 vs. gpt-3.5: A concise showdown," 2023.

[22] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," *arXiv preprint arXiv:2202.10936 , year=2022*.

[23] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, *et al.*, "Augmented language models: a survey," *arXiv preprint arXiv:2302.07842*, 2023.

[24] R. Qureshi, M. Irfan, H. Ali, A. Khan, A. S. Nittala, S. Ali, A. Shah, T. M. Gondal, F. Sadak, Z. Shah, *et al.*, "Artificial intelligence and biosensors in healthcare and its clinical relevance: A review," *IEEE Access*, 2023.

[25] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[26] M. Sallam, "The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations," *medRxiv*, pp. 2023–02, 2023.

[27] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[28] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374 , year=2021*.

[29] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "Gpts are gpts: An early look at the labor market impact potential of large language models," *arXiv preprint arXiv:2303.10130 , year=2023*.

[30] Z. Sun, "A short survey of viewing large language models in legal aspect," *arXiv preprint arXiv:2303.09136 , year=2023*.

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

[32] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[33] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, *et al.*, "Challenging big-bench tasks and whether chain-of-thought can solve them," *arXiv preprint arXiv:2210.09261*, 2022.

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

[35] Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, *et al.*, "C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models," *arXiv preprint arXiv:2305.08322*, 2023.

[36] W. Chen and E. W. X. M. J. X. T. X. X. W. P. L. Ming Yin, Max Ku, "Theoremqa: A theorem-driven question answering dataset," *arXiv preprint arXiv:2305.12524*, 2023.

[37] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[38] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[39] Y. Wang, H. Le, A. D. Gotmare, N. D. Bui, J. Li, and S. C. Hoi, "Codet5+: Open code large language models for code understanding and generation," *arXiv preprint arXiv:2305.07922*, 2023.

[40] S. Barke, M. B. James, and N. Polikarpova, "Grounded copilot: How programmers interact with code-generating models," *Proceedings of the ACM on Programming Languages*, vol. 7, no. OOPSLA1, pp. 85–111, 2023.

[41] L. Mescheder, S. Nowozin, and A. Geiger, "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks," in *International conference on machine learning*, pp. 2391–2400, PMLR, 2017.

[42] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[43] W. Zeng, X. Ren, T. Su, H. Wang, Y. Liao, Z. Wang, X. Jiang, Z. Yang, K. Wang, X. Zhang, *et al.*, "Pangu-$alpha$: Large-scale autoregressive pretrained chinese language models with auto-parallel computation," *arXiv preprint arXiv:2104.12369*, 2021.

[44] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10076–10085, 2020.

[45] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv preprint arXiv:1910.14659*, 2019.

[46] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on twitter," *Future Internet*, vol. 12, no. 11, p. 187, 2020.

[47] B. D. Lund, T. Wang, N. R. Mannuru, B. Nie, S. Shimray, and Z. Wang, "Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing," *Journal of the Association for Information Science and Technology*, vol. 74, no. 5, pp. 570–581, 2023.

[48] E. D. Liddy, "Natural language processing," 2001.

[49] X. Liu and W. B. Croft, "Statistical language modeling," *Annual Review of Information Science and Technology*, vol. 39, p. 1, 2004.

[50] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, vol. 1, p. 67, 2005.

[51] P. Azunre, *Transfer learning for natural language processing*. Simon and Schuster, 2021.

[52] A. Kovačević and D. Kečo, "Bidirectional lstm networks for abstractive text summarization," in *Advanced Technologies, Systems, and Applications VI: Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT) 2021*, pp. 281–293, Springer, 2022.

[53] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[54] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

[55] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[56] N. A. Akbar, I. Darmayanti, S. M. Fati, and A. Muneer, "Deep learning of a pre-trained language model's joke classifier using gpt-2," *Journal of Hunan University Natural Sciences*, vol. 48, no. 8, 2021.

[57] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[58] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[59]

[60] A. Karpathy, "State of GPT." https://www.youtube.com/watch?v=bZQun8Y4L2A, 2023.

[61] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," *arXiv preprint arXiv:2304.01373 , year=2023*.

[62] M. R. Chavez, T. S. Butler, P. Rekawek, H. Heo, and W. L. Kinzler, "Chat generative pre-trained transformer: why we should embrace this technology," *American Journal of Obstetrics and Gynecology*, 2023.

[63] H. Hassani and E. S. Silva, "The role of chatgpt in data science: how ai-assisted conversational interfaces are revolutionizing the field," *Big data and cognitive computing*, vol. 7, no. 2, p. 62, 2023.

[64] S. Praveen and V. Vajrobol, "Understanding the perceptions of healthcare researchers regarding chatgpt: a study based on bidirectional encoder representation from transformers (bert) sentiment analysis and topic modeling," *Annals of Biomedical Engineering*, pp. 1–3, 2023.

[65] W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, "Best: Bert pre-training for sign language recognition with coupling tokenization," *arXiv preprint arXiv:2302.05075 , year=2023*.

[66] L. Jiarong, X. Hong, J. Wenchao, Y. Jianren, and W. Tao, "Knowledge enhanced bert based on corpus associate generation," in *Machine Learning for Cyber Security: 4th International Conference, ML4CS 2022, Guangzhou, China, December 2–4, 2022, Proceedings, Part III*, pp. 533–547, Springer, 2023.

[67] M. Irfan, A. I. Sanka, Z. Ullah, and R. C. Cheung, "Reconfigurable content-addressable memory (CAM) on FPGAs: A tutorial and survey," *Future Generation Computer Systems*, vol. 128, pp. 451–465, 2022.

[68] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," *arXiv preprint arXiv:2304.02020*, 2023.

[69] J. Su, S. Yu, and D. Luo, "Enhancing aspect-based sentiment analysis with capsule network," *IEEE Access*, vol. 8, pp. 100551–100561, 2020.

[70] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[71] A. Qamar, F. B. Muslim, F. Gregoretti, L. Lavagno, and M. T. Lazarescu, "High-level synthesis for semi-global matching: Is the juice worth the squeeze?," *IEEE Access*, vol. 5, pp. 8419–8432, 2016.

[72] W. Ahmad, B. Ayrancioglu, and I. Hamzaoglu, "Low error efficient approximate adders for fpgas," *IEEE Access*, vol. 9, pp. 117232–117243, 2021.

[73] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[74] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," *arXiv preprint arXiv:1909.05858*, 2019.

[75] P. Li, M. Zhang, P. Lin, J. Wan, and M. Jiang, "Conditional embedding pre-training language model for image captioning," *Neural Processing Letters*, vol. 54, no. 6, pp. 4987–5003, 2022.

[76] I. Dergaa, K. Chamari, P. Zmijewski, and H. B. Saad, "From human writing to artificial intelligence generated text: examining the prospects and potential threats of chatgpt in academic writing," *Biology of Sport*, vol. 40, no. 2, pp. 615–622, 2023.

[77] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[78] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311 , year=2022*.

[79] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, *et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint arXiv:2112.11446 , year=2021*.

[80] D. Hernandez, T. Brown, T. Conerly, N. DasSarma, D. Drain, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, T. Henighan, T. Hume, *et al.*, "Scaling laws and interpretability of learning from repeated data," *arXiv preprint arXiv:2205.10487 , year=2022*.

[81] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," *arXiv preprint arXiv:2202.07646*, 2022.

[82] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLoS digital health*, vol. 2, no. 2, p. e0000198, 2023.

[83] M. Sallam, "Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns," in *Healthcare*, vol. 11, p. 887, MDPI, 2023.

[84] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash, *et al.*, "How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, no. 1, p. e45312, 2023.

[85] A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, and M. D. Succi, "Evaluating chatgpt as an adjunct for radiologic decision-making," *medRxiv*, pp. 2023–02, 2023.

[86] D. Duong and B. D. Solomon, "Analysis of large-language model versus human performance for genetics questions," *medRxiv*, pp. 2023–01, 2023.

[87] N. Fijačko, L. Gosak, G. Štiglic, C. T. Picard, and M. J. Douma, "Can chatgpt pass the life support exams without entering the american heart association course?," *Resuscitation*, vol. 185, 2023.

[88] R. S. D'Amico, T. G. White, H. A. Shah, and D. J. Langer, "I asked a chatgpt to write an editorial about how we can incorporate chatbots into neurosurgical research and patient care…," 2022.

[89] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, and H. Müller, "Ai for life: Trends in artificial intelligence for biotechnology," *New Biotechnology*, vol. 74, pp. 16–24, 2023.

[90] M. R. Haque and S. Rubya, "An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews," *JMIR mHealth and uHealth*, vol. 11, no. 1, p. e44838, 2023.

[91] S. A. Basit, R. Qureshi, S. Musleh, R. Guler, M. S. Rahman, K. H. Biswas, and T. Alam, "Covid-19base v3: Update of the knowledgebase

for drugs and biomedical entities linked to covid-19," *Frontiers in Public Health*, vol. 11, p. 1125917, 2023.

[92] mbzuai oryx, "Xraygpt: Chest radiographs summarization using medical vision-language models," 2023.

[93] J. S. () and W. Y. (), "Unlocking the power of chatgpt: A framework for applying generative ai in education," *ECNU Review of Education*, vol. 0, no. 0, p. 20965311231168423, 0.

[94] "An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 2, no. 4, p. 100089, 2022.

[95] H. Crompton and D. Burke, "Artificial intelligence in higher education: the state of the field," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 22, 2023.

[96] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[97] "Khan academy explores the potential for gpt-4 in a limited pilot program," 2023.

[98] "Harnessing gpt-4 so that all students benefit. a nonprofit approach for equal access," 2023.

[99] E. Hannan and S. Liu, "Ai: new source of competitiveness in higher education," *Competitiveness Review: An International Business Journal*, vol. 33, no. 2, pp. 265–279, 2023.

[100] Z. Lin, Z. Song, Z. Dai, and Q. V. Le, "Fingpt: Open-source financial large language models," *arXiv preprint arXiv:2306.06031*, 2023.

[101] M. Fraiwan and N. Khasawneh, "A review of chatgpt applications in education, marketing, software engineering, and healthcare: Benefits, drawbacks, and research directions," *arXiv preprint arXiv:2305.00237*, 2023.

[102] D. Tiro, "The possibility of applying chatgpt (ai) for calculations in mechanical engineering," in *New Technologies, Development and Application VI: Volume 1*, pp. 313–320, Springer, 2023.

[103] Y. Wardat, M. A. Tashtoush, R. AlAli, and A. M. Jarrah, "Chatgpt: A revolutionary tool for teaching and learning mathematics," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 19, no. 7, p. em2286, 2023.

[104] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner, "Mathematical capabilities of chatgpt," *arXiv preprint arXiv:2301.13867 , year=2023*.

[105] X. Wang, N. Anwer, Y. Dai, and A. Liu, "Chatgpt for design, manufacturing, and education," 2023.

[106] S. Badini, S. Regondi, E. Frontoni, and R. Pugliese, "Assessing the capabilities of chatgpt to improve additive manufacturing troubleshooting," *Advanced Industrial and Engineering Polymer Research*, 2023.

[107] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, p. 100006, 2020.

[108] J. Rudolph, S. Tan, and S. Tan, "War of the chatbots: Bard, bing chat, chatgpt, ernie and beyond. the new ai gold rush and its impact on higher education," *Journal of Applied Learning and Teaching*, vol. 6, no. 1, 2023.

[109] T. Reddy, R. Williams, and C. Breazeal, "Text classification for ai education.," in *SIGCSE*, p. 1381, 2021.

[110] J. Pachouly, S. Ahirrao, K. Kotecha, G. Selvachandran, and A. Abraham, "A systematic literature review on software defect prediction using artificial intelligence: Datasets, data validation methods, approaches, and tools," *Engineering Applications of Artificial Intelligence*, vol. 111, p. 104773, 2022.

[111] K. Nguyen-Trung, A. K. Saeri, and S. Kaufman, "Applying chatgpt and ai-powered tools to accelerate evidence reviews," 2023.

[112] N. Gleason, "Chatgpt and the rise of ai writers: How should higher education respond?," *Times Higher Education*, 2022.

[113] G. Cooper, "Examining science education in chatgpt: An exploratory study of generative artificial intelligence," *Journal of Science Education and Technology*, vol. 32, pp. 444–452, 2023.

[114] L. Skavronskaya, A. H. Hadinejad, and D. Cotterell, "Reversing the threat of artificial intelligence to opportunity: a discussion of chatgpt in tourism education," *Journal of Teaching in Travel & Tourism*, vol. 23, no. 2, pp. 253–258, 2023.

[115] B. Yetiştiren, I. Özsoy, M. Ayerdem, and E. Tüzün, "Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, amazon codewhisperer, and chatgpt," *arXiv preprint arXiv:2304.10778*, 2023.

[116] A. M. Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais, and Z. M. J. Jiang, "Github copilot ai pair programmer: Asset or liability?," *Journal of Systems and Software*, vol. 203, p. 111734, 2023.

[117] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the era of chatgpt et al. the prospects of large language models," *Business & Information Systems Engineering*, pp. 1–7, 2023.

[118] C. Xu, Y. Xu, S. Wang, Y. Liu, C. Zhu, and J. McAuley, "Small models are valuable plug-ins for large language models," *arXiv preprint arXiv:2305.08848*, 2023.

[119] J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J.-R. Wen, "Recommendation as instruction following: A large language model empowered recommendation approach," *arXiv preprint arXiv:2305.07001*, 2023.

[120] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, *et al.*, "Summary of chatgpt/gpt-4 research and perspective towards the future of large language models," *arXiv preprint arXiv:2304.01852*, 2023.

[121] X. He, X. Shen, Z. Chen, M. Backes, and Y. Zhang, "Mgtbench: Benchmarking machine-generated text detection," *arXiv preprint arXiv:2303.14822*, 2023.

[122] M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp," *arXiv preprint arXiv:2305.14976*, 2023.

[123] J. Kim, J. H. Lee, S. Kim, J. Park, K. M. Yoo, S. J. Kwon, and D. Lee, "Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization," *arXiv preprint arXiv:2305.14152*, 2023.

[124] S. Arora, B. Yang, S. Eyuboglu, A. Narayan, A. Hojel, I. Trummer, and C. Ré, "Language models enable simple systems for generating structured views of heterogeneous data lakes," *arXiv preprint arXiv:2304.09433*, 2023.

[125] S. R. Bowman, "Eight things to know about large language models," *arXiv preprint arXiv:2304.00612*, 2023.

[126] M. J. Ali, "Chatgpt and lacrimal drainage disorders: performance and scope of improvement," *Ophthalmic Plastic and Reconstructive Surgery*, vol. 39, no. 3, p. 221, 2023.

[127] Y. Fu, H. Peng, T. Khot, and M. Lapata, "Improving language model negotiation with self-play and in-context learning from ai feedback," *arXiv preprint arXiv:2305.10142*, 2023.

[128] P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, 2023.

[129] H. Matsumi, D. Hallinan, D. Dimitrova, E. Kosta, and P. De Hert, *Data Protection and Privacy, Volume 15: In Transitional Times*. Bloomsbury Publishing, 2023.

[130] P. Hacker, A. Engel, and M. Mauer, "Regulating chatgpt and other large generative ai models," *arXiv preprint arXiv:2302.02337*, 2023.

[131] S. A. Khowaja, P. Khuwaja, and K. Dev, "Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review," *arXiv preprint arXiv:2305.03123*, 2023.

[132] A. Chan, H. Bradley, and N. Rajkumar, "Reclaiming the digital commons: A public data trust for training data," *arXiv preprint arXiv:2303.09001*, 2023.

[133] W. H. Deng, B. Guo, A. Devrio, H. Shen, M. Eslami, and K. Holstein, "Understanding practices, challenges, and opportunities for user-engaged algorithm auditing in industry practice," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2023.

[134] M. Kraus, J. A. Bingler, M. Leippold, T. Schimanski, C. C. Senni, D. Stammbach, S. A. Vaghefi, and N. Webersinke, "Enhancing large language models with climate resources," *arXiv preprint arXiv:2304.00116*, 2023.

[135] E. Agathokleous, C. J. Saitanis, C. Fang, and Z. Yu, "Use of chatgpt: What does it mean for biology and environmental science?," *Science of The Total Environment*, p. 164154, 2023.

[136] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, and L. Moy, "Chatgpt and other large language models are double-edged swords," 2023.

[137] Z. Sun, Y. Shen, Q. Zhou, H. Zhang, Z. Chen, D. Cox, Y. Yang, and C. Gan, "Principle-driven self-alignment of language models from scratch with minimal human supervision," *arXiv preprint arXiv:2305.03047*, 2023.

[138] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," *arXiv preprint arXiv:2304.03738*, 2023.

[139] Y. Wolf, N. Wies, Y. Levine, and A. Shashua, "Fundamental limitations of alignment in large language models," *arXiv preprint arXiv:2304.11082*, 2023.
[140] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting llm-generated texts," *arXiv preprint arXiv:2303.07205*, 2023.
[141] F. Ufuk, "The role and limitations of large language models such as chatgpt in clinical settings and medical journalism," *Radiology*, vol. 307, no. 3, p. e230276, 2023.
[142] R. Bhayana, S. Krishna, and R. R. Bleakney, "Performance of chatgpt on a radiology board-style examination: Insights into current strengths and limitations," *Radiology*, p. 230582, 2023.
[143] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?," *arXiv preprint arXiv:2305.01937*, 2023.
[144] X. Yang, Y. Li, X. Zhang, H. Chen, and W. Cheng, "Exploring the limits of chatgpt for query or aspect-based text summarization," *arXiv preprint arXiv:2302.08081*, 2023.
[145] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 258–268, 2022.
[146] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3428–3448, Association for Computational Linguistics, July 2019.
[147] J. Weston, E. Dinan, and A. Miller, "Retrieve and refine: Improved sequence generation models for dialogue," in *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, (Brussels, Belgium), pp. 87–92, Association for Computational Linguistics, Oct. 2018.
[148] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 809–819, Association for Computational Linguistics, June 2018.
[149] D. V. Hada and S. K. Shevade, "Rexplug: Explainable recommendation using plug-and-play language model," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 81–91, 2021.
[150] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-rec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524 , year=2023*.
[151] A. Uchendu, *REVERSE TURING TEST IN THE AGE OF DEEPFAKE TEXTS*. PhD thesis, The Pennsylvania State University, 2023.
[152] E. M. Bonsu and D. Baffour-Koduah, "From the consumers' side: Determining students' perception and intention to use chatgpt in ghanaian higher education," *Journal of Education, Society & Multiculturalism*, vol. 4, no. 1, pp. 1–29, 2023.
[153] Q. V. Liao and J. W. Vaughan, "AI Transparency in the Age of LLMs: A human-centered research roadmap," *arXiv preprint arXiv:2306.01941 , year=2023*.
[154] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.
[155] N. M. Deshpande, S. Gite, B. Pradhan, and M. E. Assiri, "Explainable artificial intelligence–a new step towards the trust in medical diagnosis with ai frameworks: A review," *Comput. Model. Eng. Sci*, vol. 133, pp. 1–30, 2022.
[156] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International Journal of Human-Computer Studies*, vol. 146, p. 102551, 2021.
[157] M. S. Rahaman, M. T. Ahsan, N. Anjum, H. J. R. Terano, and M. M. Rahman, "From chatgpt-3 to gpt-4: A significant advancement in ai-driven nlp tools," *Journal of Engineering and Emerging Technologies*, vol. 2, no. 1, pp. 1–11, 2023.
[158] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, "Risks and benefits of large language models for the environment," *Environmental Science & Technology*, vol. 57, no. 9, pp. 3464–3466, 2023.
[159] G. Fergusson, C. Fitzgerald, C. Frascella, M. Iorio, T. McBrien, C. Schroeder, B. Winters, and E. Zhou, "Contributions by,"
[160] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," 2021.
[161] A. S. George, A. H. George, and A. G. Martin, "The environmental impact of ai: A case study of water consumption by chat gpt," *Partners Universal International Innovation Journal*, vol. 1, no. 2, pp. 97–104, 2023.
[162] S. Biswas, "Potential use of chat gpt in global warming," *Ann Biomed Eng*, vol. 51, pp. 1126–1127, 2023.
[163] Z. Yao, Y. Lum, A. Johnston, and et al., "Machine learning for a sustainable energy future," *Nat Rev Mater*, vol. 8, pp. 202–215, 2023.
[164] X. Zhi and J. Wang, "Editorial: Ai-based prediction of high-impact weather and climate extremes under global warming: A perspective from the large-scale circulations and teleconnections," *Frontiers in Earth Science*, vol. 11, 2023.
[165] J. Zhong, Y. Zhong, M. Han, T. Yang, and Q. Zhang, "The impact of ai on carbon emissions: evidence from 66 countries," *Applied Economics*, vol. 0, no. 0, pp. 1–15, 2023.
[166] M. A. Habila, M. Ouladsmane, and Z. A. Alothman, "Chapter 21 - role of artificial intelligence in environmental sustainability," in *Visualization Techniques for Climate Change with Machine Learning and Artificial Intelligence* (A. Srivastav, A. Dubey, A. Kumar, S. Kumar Narang, and M. Ali Khan, eds.), pp. 449–469, Elsevier, 2023.
[167] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, pp. 433–460, 1950.
[168] F. O. Letters, "Pause giant ai experiments: An open letter," *Future of Life Institution. https://futureoflife. org/open-letter/pause-giant-ai-experiments*, 2023.
[169] M. Ienca, "Don't pause giant ai for the wrong reasons," *Nature Machine Intelligence*, pp. 1–2, 2023.
[170] B. Lin, D. Bouneffouf, G. Cecchi, and K. R. Varshney, "Towards healthy ai: Large language models need therapists too," *arXiv preprint arXiv:2304.00416 , year=2023*.
[171] S. Harrer, "Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine," *EBioMedicine*, vol. 90, 2023.
[172] M. Elmahdy and R. Sebro, "A snapshot of artificial intelligence research 2019–2021: is it replacing or assisting physicians?," *Journal of the American Medical Informatics Association*, p. ocad094, 2023.
[173] C. T. Wolf, "Democratizing ai? experience and accessibility in the age of artificial intelligence," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 26, no. 4, pp. 12–15, 2020.