Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects

Muhammad Usman Hadi [1], qasem al tashi [1], Rizwan Qureshi [2,2,2,2], Abbas Shah [1], amgad muneer [1], Muhammad Irfan [1], Anas Zafar [1], Muhammad Bilal Shaikh [1], Naveed Akhtar [1], Jia Wu [1], and Seyedali Mirjalili [1]

[1]Affiliation not available
[2]MD Anderson Cancer Center

December 7, 2023

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

Large Language Models: A Comprehensive Survey of Applications, Challenges, Limitations, and Future Prospects

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], Mohammed Ali Al-Garadi[8], Jia Wu[2], and Seyedali Mirjalili[9,10]

[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 (zaigham.shah@faculty.muet.edu.pk)*
[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]*Computing and Information Systems, The University of Melbourne, 700 Swanston Street, Carlton 3010, VIC Australia*
[8]*Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA*
[9]*Centre for Artificial Intelligence Research and Optimization, Torrens University Australia, Fortitude Valley, Brisbane, QLD 4006, Australia (ali.mirjalili@torrens.edu.au)*
[10]*University Research and Innovation Center, Obuda University, 1034 Budapest, Hungary*
** Corresponding Author*

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. 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, vision applications, and question-answering. This survey provides a comprehensive overview of LLMs, including their history, architecture, training methods, applications, and challenges. We begin by discussing the fundamental concepts of generative AI and the architecture of generative pre-trained transformers (GPT). We then provide an overview of the history of LLMs, their evolution over time, and the different training methods that have been used to train them. We then discuss the wide range of tasks where they are used and also discuss applications of LLMs in different domains, including medicine, education, finance, engineering, media, entertainment, politics, and law. We also discuss how LLMs are shaping the future of AI and their increasing role in scientific discovery, and how they can be used to solve real-world problems. Next, we explore the challenges associated with deploying LLMs in real-world scenarios, including ethical considerations, model biases, interpretability, and computational resource requirements. This survey also highlights techniques for enhancing the robustness and controllability of LLMs and addressing bias, fairness, and quality issues in Generative AI. Finally, we conclude by highlighting the future of LLM research and the challenges that need to be addressed in order to make this technology more reliable and useful. This survey 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 article is anticipated to serve as a valuable resource for learning the current state-of-the-art as well as 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 Github-Repo.

Index Terms

Large Language Models, Large Vision Models, Generative AI, Conversational AI, LangChain, Natural language processing, Computer Vision, GPT, ChatGPT, Bard, AI chatbots

Large Language Models: A Comprehensive Survey of Applications, Challenges, Limitations, and Future Prospects

I. Introduction

Language modeling (LM) is a fundamental task in natural language processing (NLP) that aims to predict the next word or a character in a given sequence of text [1], [2]. It involves developing algorithms and models that can understand and generate coherent human language. The primary objective of LM is to capture the probability distribution of words in a language, which allows the model to generate new text [3], complete sentences [4], and predict the likelihood of different word sequences [5], [6]. They are broadly categorized into statistical language models, machine learning models, deep learning models, and transformer based models as shown in Fig. 1. Early language models, such as n-gram models [7], were based on simple statistical techniques that estimated the probabilities of word sequences using frequency counts [8], [9]. However, with the rise of deep learning in NLP [10], the availability of enormous amounts of public datasets [11], and powerful computing devices [12] to process these big data with complex algorithms, has led to the development of large language models.

Large Language Models (LLMs) [13], sometimes referred to as "transformative [14]" or "next-generation [15]" language models, represent a significant breakthrough in NLP [16]. These models leverage deep learning techniques, particularly transformer architectures [17], to learn and understand the complex patterns and structures present in language data [18]. A key characteristic of LLMs is their ability to process vast amounts of data, including unstructured text, and capture semantic relationships between words and phrases [19]. These models can also process visual [20], audio [21], audiovisual [22], as well as multi-modal data [23] and learn the semantic relationships between them. These models have significantly enhanced the capabilities of machines to understand and generate human-like language [24].

The history of LLMs can be traced back to the early development of language models and neural networks [25]. The journey begins with the era of statistical language models [26]. In this stage, researchers primarily relied on probabilistic approaches [27] to predict word sequences. Classic examples include n-grams, Hidden Markov Models (HMMs) [28] and Maximum Entropy Models [29]. N-grams, for instance, are sequences of adjacent words or tokens that are used to predict the likelihood of the next word based on the preceding ones [30]. While rudimentary by today's standards, these models marked a crucial starting point in the field of natural language understanding. They allowed for basic text generation and word prediction but were limited in their ability to capture complex contextual relationships [31] [32]. [33]. Then a shift towards more data-driven methodologies has been witnessed [34]. Researchers began to explore machine learning algorithms to improve language understanding [35] These models learned patterns and relationships within large text corpora. Support Vector Machines (SVMs) is a notable example from this [36]. Machine learning models brought a more sophisticated approach to NLP tasks, allowing for the development of applications like spam detection [37] and sentiment analysis [38]. Moreover the availability of large-scale Twitter[1] datasets has brought a revolution in real time sentiment analysis [39].

The emergence of deep learning marked a pivotal moment in the development of LLMs [40]. Neural networks, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks gained prominence [41]. These deep learning architectures delved deeper into the data, allowing them to capture more intricate features and long-range dependencies within text. This stage significantly improved the models' ability to understand context, making them suitable for tasks like machine translation and speech recognition [16], [42]. However, deep learning also faced challenges with vanishing gradients [43] and long-term dependencies [44], limiting their effectiveness.

The breakthrough in LLMs came with the introduction of the Transformer architecture in the seminal work "Attention is All You Need" by Vaswani et al. in 2017 [45]. The Transformer model, based on the self-attention mechanism [46], enabled parallelization and efficient handling of long-range dependencies. It laid the foundation for models like OpenAI's GPT (Generative Pre-trained Transformer) series [47] and BERT (Bidirectional Encoder Representations from Transformers) [48] by Google, which achieved groundbreaking results in a wide range of language tasks. These mechanisms enabled models to consider the entire context of a sentence or document, allowing for true contextual understanding [49]. Transformer-based models, often pre-trained on massive text corpora, can generate coherent and contextually relevant text, revolutionizing applications like chatbots [50], text summarization [51], and language translation [52].

ChatGPT, Llama, and Falcon are all remarkable variants of the GPT (Generative Pre-trained Transformer) model [53], which is developed and pioneered by OpenAI. These models represent OpenAI's ongoing efforts to push the boundaries of

[1]Twitter, as a microblogging platform, allows users to express their thoughts and opinions in short, concise messages called tweets. These tweets often contain rich, real-time information about various topics, making Twitter an excellent source for sentiment analysis

natural language processing and understanding. They share a common foundation in their training methodology, which involves pre-training on vast corpora of text data followed by fine-tuning for specific tasks. During pre-training, the models are exposed to diverse internet text to learn grammar, facts, reasoning abilities, and some degree of common-sense knowledge [54], [55]. This process equips them with a broad understanding of language. Subsequently, fine-tuning is carried out on narrower datasets to specialize the models for particular applications. ChatGPT, for instance, is fine-tuned for conversational contexts, making it suitable for chatbots and virtual assistants [56], [57], [58]. Llama and Falcon, though not as widely known, represent potential advancements or specialized versions, possibly tailored for specific use cases or research objectives. These models collectively exemplify the cutting-edge advancements in natural language processing, enabling more human-like interactions and understanding through the power of AI-driven language models [59], [60], [61].

The training process for models like ChatGPT, Llama, and Falcon consists of several key stages [62], [53]. It begins with a phase known as pre-training, where these models are exposed to a vast and varied dataset of internet text, enabling them to learn grammar, vocabulary, world knowledge, and context [63]. The underlying architecture, based on the Transformer model, is crucial for understanding the relationships between words in sentences. Following pre-training, the models undergo fine-tuning on specific datasets tailored to particular tasks, such as text generation or conversation in the case of ChatGPT. Fine-tuning refines their capabilities for these specialized tasks, with hyperparameter tuning to optimize performance. Ethical considerations are an integral part of the process, aiming to minimize harmful or biased outputs. It's important to note that this training is an iterative and resource-intensive endeavor, continually improved and monitored to enhance both performance and safety [64], [65].

LLMs have undergone several developmental stages, with models increasing in size and complexity. The GPT series, starting with GPT-1 and continuing with GPT-2 and GPT-3 [66], has successively grown in the number of parameters, starting from hundreds of millions (GPT-1) to 1.7 Trillion (GPT-4) [67], allowing for more sophisticated language understanding and generation capabilities [68]. Similarly, BERT-inspired models have seen advancements in pre-training strategies, such as ALBERT [69] (A Lite BERT) and RoBERTa [70], which further improved performance and efficiency.

Furthermore, advancements in LLMs have extended to more specific domains, with models designed for specialized tasks like medical language processing [71], scientific research [72], website development [73] and code generation [74]. Moreover, efforts have been made to address ethical concerns [75], interpretability [76], and reducing biases in LLMs to ensure responsible and equitable use [77]. The development stages of large models have witnessed a constant quest for larger models, improved pre-training strategies, and specialized domain adaptations [78], [79]. As research continues, the potential applications and impact of LLMs on various fields, including education, healthcare, and human-computer interaction, continue to expand, inspiring further innovations and advancements.

TABLE I: List of Acronyms and corresponding definitions.

Acronym	Definition
AI	Artificial Intelligence
AGI	Artificial General Intelligence
BBH	Big Bench Hard
BERT	Bidirectional Encoder Representations from Transformers
CV	Computer Vision
ChatGPT	A Large Language Model by OpenAI
CTRL	Conditional Transformer Language Model
FFF	Fused Filament Fabrication
GANs	Generative Adversarial Networks
GNMT	Google Neural Machine Translation
GPT	Generative Pre-Trained transformers
GenAI	Generative AI
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
RLHF	Reinforcement Learning Human Feedback
RNN	Recurrent neural networks
RNNLM	Recurrent neural network language model
SLMs	Statistical Language Models
T2V	Text to video
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

In summary and as can be seen from Fig 1; LM research has received widespread attention and has undergone four significant development stages including: statistical language models, machine learning models, deep learning models and transformer-based models[2]. In this research, we mainly focus on LLMs and foundation AI models for language and vision tasks. A list of commonly used acronyms in this article with definitions is given in Table I.

Modern language model called ChatGPT [80] was developed by OpenAI [81] and launched in 2022. It is based on the GPT-3.5 architecture [82] and was trained using a sizable amount of internet-sourced text data, including books, articles, wikis and websites (Table II) [83]. ChatGPT is exceptional at producing human-like responses and having conversations with users.

In computer vision (CV), researchers are also actively engaged in the development of vision-language models inspired by the capabilities of ChatGPT. These models are specifically designed to enhance multimodal dialogues, where both visual and textual information are important [84]. Moreover, the advancements in the field have led to the introduction of GPT-4 [82], which has further expanded the capabilities of language models by seamlessly integrating visual information

[2]Due to the dominance of Transformer based models, we consider it a different stage, not as a subset of deep learning

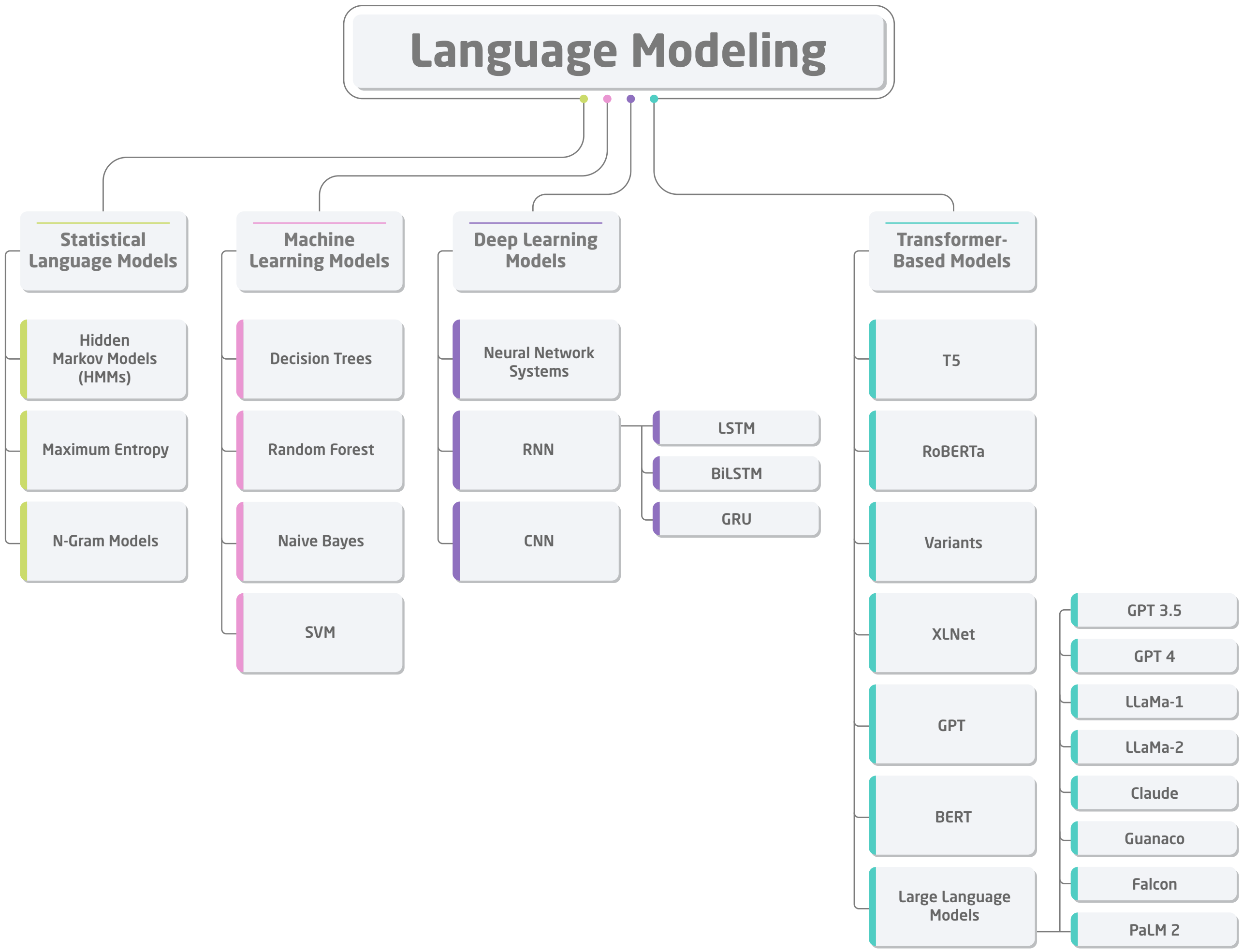


Fig. 1: Types of language modeling. The division of LLMs is categorized into four major blocks: Statistical language models, Machine learning models, Deep learning models and Transformer-based models.

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 [85], [86], [87], [88], [89], [90], [91]. For example, the role of ChatGPT in education is evaluated in [92], healthcare and medicine in [93], [71], protein sequence modeling in [94] and protein generation in [95]. A survey on generative AI is presented in [96], scientific text modeling in [97] and text generation in [98]. The use of LLM in finance is evaluated in [99], impact on labor market in [100] and supply chain in [101], telecom in [102], on code writing capabilities in [103], deep fakes in [104], legal aspects in [105], AI for drug discovery in [106], clinical prediction with LLM in [107], ML for cancer biomarkers in [108], and integration of biotechnology and AI applications to address global challenges in [109]. The advancements in pre-training, fine-tuning, utilization and capability evaluation of LLMs is presented in [85] and a survey on autonomous agents in [110]. The recent progress in visio-language pre-trained models is discussed in [86] and the knowledge graphs construction and reasoning are explained in [111], selection inference is in [76], and quantum-inspired machine learning in [112]. The survey vision language pre-trained models [86] presents an overview of various techniques for encoding raw images and texts into single-modal embeddings as a fundamental aspect, and also discusses 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. Also, the field is progressing at a very fast pace, so a review article with practical applications will contribute a lot to the field. Therefore we also write this paper in the form of an application oriented review. 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 others, and discuss strategies for their optimal utilization. Furthermore, we provide detailed insights about generative AI, writing prompts, and visual prompting techniques, leveraging GPT-plug-ins, and harnessing other AI/LLM tools. These aspects are generally not covered by the existing related articles. Our comprehensive examination also encompasses a discussion on the limitations associated with the 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 paper will contribute to a better understanding and utilization of LLMs.

B. Contributions

The main contributions of this article are as follows:

1) Providing a comprehensive overview of GenAI and 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. We also discuss the limitations of LLMs; including, limited understanding of the physical world, tokenization problems, infomration hallucination, fine-tuning and risk of foundation models.
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, law, politics, media, entertainment, engineering, and others.
5) This article is uniquely presented in a manner to promote practical usage of LLMs, showcasing the actual LLM outputs to corroborate the discussions.

The paper is organized as the following sections. Section II provides an introduction to the role of generative AI, specifically focusing on the fundamentals of new data generation and variance, as well as its applications. Section III presents an overview of LLMs, summarizing a brief history of LLMs and discussing their training and functionality. Taxonomy of LLMs is presented in Section IV and major applications of LLM through different use cases is discussed in Section V. Section VI explores AI-enabled tools that are expected to shape the future. Section VII discusses the practical use cases of GPT plugins and their potential to enhance user productivity and efficiency. Section VIII presents guidelines and working examples using prompting techniques. Section IX proposes the limitations and drawbacks of the current state-of-the-art LLM. Section X presents open questions on the subject matter and the authors' perspective on open unanswered avenues. Section XI concludes the survey paper. The overall structure of the article is presented in Fig. 2 for a quick reference at a glance.

TABLE II: **Pre-training data**. Mixtures of data used for pre-training LLaMA [15].

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
Stock Exchange	2.0%	1.03	78GB

II. GENERATIVE AI

Generative AI (GenAI) [113] is perhaps the most disruptive [114] and generalized technology of this decade [115], already influenced many industries, including, Media [116], Marketing [117], Game development and Metaverse [118], Education [119], Software development [120], and Medical [121], construction technology [122], and pharmaceuticals [123]. Unlike general AI systems that perform specific tasks such as data classification [124], clustering [125], object detection [126] and segmentation [127] or predictions [128]; GenAI can generate meaningful new content of multiple data modalities [129]; including, text [3], speech [130], images [131], and videos [132]. Some common examples of GenAI systems are image generators (Midjourney or stable diffusion), Chatbots (ChatGPT, Bard, Palm), code generators (CodeX, Co-Pilot [133]) audio generators(VALL-E)Valle [134], and video generators (Gen-2) [135]

During the past few years, GenAI models size has been scaled from a few million parameters(BERT [48], 110M) to hundreds of billions of parameters (GPT [136], 175B). Generally speaking, as the size of the model (number of parameters) increases, the performance of the model also increases [137], and it can be generalized for a variety of tasks [138], for example, Foundation models [139]. However, smaller models can also be fine-tuned for a more focused task [140].

LLMs, such as ChatGPT by OpenAI, Bard by Google, and Llama by Meta, are a type of GenAI models, specifically designed to generate human-like language in response to a given prompt [141]. These models are trained on massive amounts of data (see Table II), using techniques 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" [142].

GenAI works by leveraging complex algorithms and statistical models to generate new content that mimics the patterns and characteristics of the training data [143]. These algorithms may include probabilistic techniques; such as Autoregressive model [144] and Variations Auto-encoders [145], or more recently, Generative Adversarial Networks [146] and Diffusion models [147] or Reinforcement Learning Human Feedback (RLHF) [148].

GenAI has captured significant interest in recent years due to its remarkable performance across an extensive array

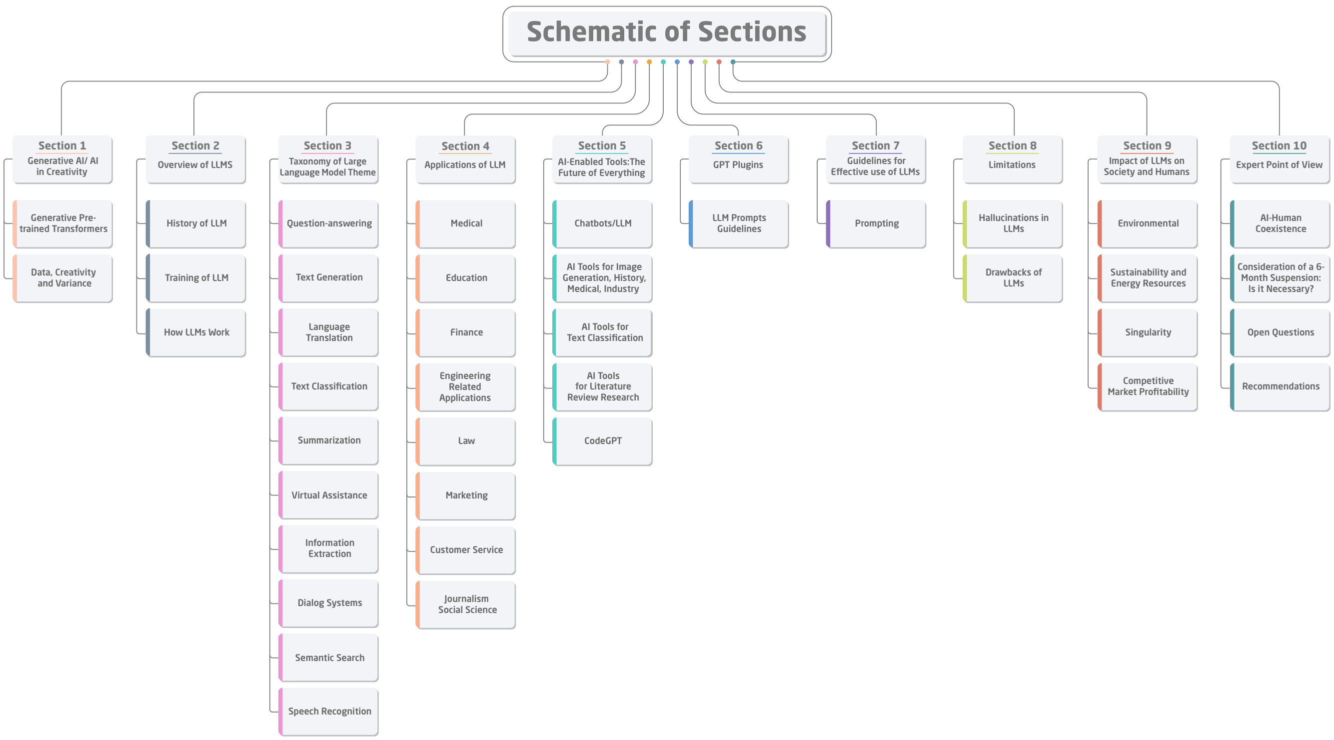


Fig. 2: Schematic Representation of Article Sections - A visual overview of the key sections comprising the structure of the article, providing readers with a roadmap for navigating the content effectively.

of applications in text, image, video and generation [149]. Constructed upon the foundation of the transformer architecture [45], these models exhibit an extraordinary capacity to process and generate human-like content by leveraging massive volumes of training data for various topics [150].

A. Data, Generation, Variance, and Performance measures

To comprehend the intricacies of GenAI systems, it is important to delve into the concepts of data, generation, and variance, and the interplay between them, as they form the foundation of generative systems [151].

1) Data: The core of generative AI systems is data. Training models that can successfully capture the underlying patterns and structures of the target domain require high-quality and diverse training data. The generating performance is influenced by the amount, quality, and representation of the training data [152], [153]. Furthermore, the availability of large-scale, labeled datasets allows for the development of more accurate and coherent samples [154], whereas restricted or biased training data may yield sub-optimal results [155].

2) Generation Process: GenAI uses the gained knowledge from the training data to generate samples with similar statistical patterns [156]. The generative models are designed to capture the underlying distributions of the training data and generate reliable and realistic samples with properties consistent with the original dataset [157]. The generating process involves approaches; such as adversarial training [158], latent space interpolation [159], and autoregressive modeling [160].

3) Variance: Variance is another important factor in defining the diversity and quality of generated samples [161], which shows the variability in the generated samples. A low variance generative AI system may produce similar or repetitive samples, resulting in poor generation, whereas, a high variance, may yield diversified but unrealistic or incoherent samples [162]. Striking a balance between variation and fidelity is difficult in generative AI [163], since it requires managing the trade-off between exploring and exploiting the learnt data distribution [164], [165].

Understanding and regulating the relationship between data, variation, and generation process, is essential for the development of efficient GenAI systems [166]. It entails dealing with issues; including dataset biases [167], mode collapse [168], and balancing exploration and exploitation [**?**]. GenAI systems may generate high-quality, diversified, and realistic samples that correspond with the desired aims and applications by refining the training data, optimizing the generation procedures, and regulating variation [169].[3]

4) Performance Metrics: Evaluating the quality and diversity of generated samples is critical for assessing the models' performance [170]. Several strategies for assessing the quality, diversity, and authenticity of generated samples have been established. Here are some common evaluation techniques:

- Visual inspection [171], which is a subjective a evaluation method where human experts or users examine the generated samples and provide qualitative feedback.

[3]genai meets copyright science paper

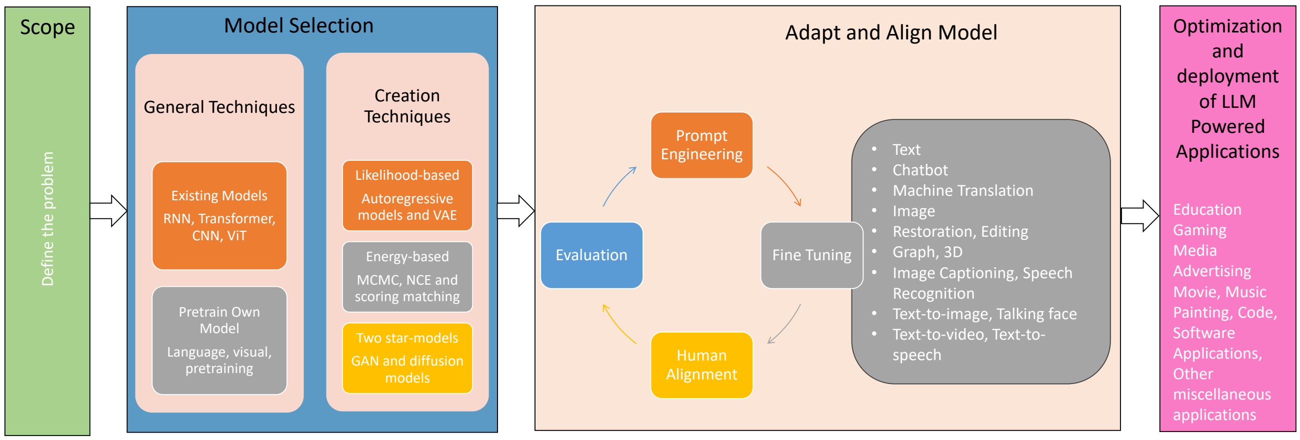


Fig. 3: Generative AI Design Process. Most of the existing generative models can be fine-tuned for any downstream task. First, you need to define the problem, and whether you can use any existing model or design a new model. Prompt engineering, where we can inject a series of prompts with the desired outcomes, fine-tuning, or human alignment can be used, to tune the model, for a specific task. Generative AI models can optimized for a variety of tasks, including, education, healthcare, entertainment and others.

- Inception Score (IS) [172] a widely used quantitative evaluation metric for, which measures the quality of generated samples based on their visual appeal and the diversity of the generated classes. Higher IS scores indicate better quality and diversity of the generated samples.
- Frechet Inception Distance (FID) [173] compares the distributions of real and generated samples by calculating the Fréchet distance between their feature representations extracted from a pre-trained Inception model.

Other performance metrics for GenAI systems, include PR-curves [174], coverage metrics [175], user studies and others [176].

B. Generative AI Design Cycle

A typical design cycle of Generative AI is shown in Figure. 3 as adopted from [129], [177]. GenAI development cycle may be broken into four key steps; (i) define the problem, (ii) model selection or developing from scratch (iii) adapt and align the model, or fine-tuning if necessary, and (iv) finally the deployment and optimization stage [178].

The first stage, scope entails deciding the the nature of the target GenAI model. For e.g., is the target to make it perform well at multiple tasks or only a single task? The nexts stage is the selection of a model. In this stage, a GenAI model developer needs to decide whether to use an existing model for the application or to pre-train one from scratch [179]. In this step, the developer can go for general techniques, for e.g. RNNs, transformers, or pretraining their own model and/or delve deep into creation methods involving more nuanced modifications of the model being used [180].

Following the second stage is an iterative phase of aligning and adapting the model for the scope chosen [181]. This includes steps of prompt engineering [79], which may consist of zero-shot learning, one-shot learning, or a few show learning techniques [182], or even fine-tuning based on the scope [183], and evaluating the model performance.

The last stage is the optimization and deployment for the target application [184]. Since prompting is a fundamental aspect of using and developing GenAI models, we provide a detailed discussion on prompting techniques in this section.

1) Prompting: LLMs have given rise to whats called "Prompt Engineering". Prompts are the instructions provided to an LLM to make it follow specified rules, automation of processes and to ensure that the output generated is of a specific quality or quantity [79], [185]. 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 therefore very important if one needs to use LLMs to assist with tasks in the best manner possible [186].

While some formal techniques such as Explicit instruction (providing a clear direction to the LLM to do something) [187], 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) [188] and Interaction and iteration/chaining (interact with model iteratively to reach to a good answer by fine-tuning on each reply) have been presented [79].

Several different frameworks have been suggested in lieu of prompt patterns for LLMs, these are generic prompt patterns targetting a specific category such as prompt improvement, input semantics etc [79], [189], [190], or prompting for software engineering tasks [191], [192], however, in this work, we aim to present some sets of commands to help users get the most out of the LLMs capabilities from a generic perspective.

- *Defining the role/context*: 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 [193]. A similar first prompt could be providing the context. This can be performed to give the LLM a background of the conditions in which the LLM is supposed to work. For e.g., "We are a company performing mobile application development for Fortune 500 organizations". This can then be followed up with aspects like actions, tasks to perform, steps to follow, etc as mentioned before [194].
- *Prompt creation*: Another interesting prompt command is to ask the model to generate prompts for a certain task [195]. 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”.
- *Chain of thoughts*: Chain of thoughts prompting [196] in the context of Language Models (LMs) refers to the practice of providing a series of related prompts or partial sentences to guide the generation of coherent and connected text. Instead of providing a single prompt, a chain of thoughts prompt involves providing multiple prompts in succession to encourage the LM to continue generating text that follows a specific line of thinking or narrative [197].
- Other interesting directions in which Prompts can be given are explanation prompts [198] (e.g., “Explain 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”) [199].

One concept within prompt engineering is in-context learning [200] in terms of the user “teaching” the LLM to act in a certain manner. The typical prompting scheme in which the LLM is asked to perform a task is an example of zero-shot inference [201], that is, within the context of the current task being worked on, the LLM is asked to perform the task without providing any sample solution for it. An example of this type prompt could happen in the task of classifying tweets. To perform zero-shot inference, a user will have to just provide the text of the tweet to the LLM and ask it to classify it as positive or negative in sentiment. Another type of prompting could be one-shot inference [202]. In such a case, the user would give an example of a task solution o the LLM and then ask it to perform the task. In the tweet sentiment analysis example previously, this would be the user providing a sample of a tweet and information the LLM that the sentiment is positive and then providing it a second tweet to determine the sentiment of. The third type of prompt is few-shot inference [203], herein, the user provides a few examples of task solutions to teach the LLM about the kind of operation the user wants it to do. For the tweet sentiment analysis example above, it would be providing a tweet/tweets with a positive sentiment and indicating its sentiment and doing so with a negative tweet/tweets as well. Finally, the user can then use the LLM for tweet classification. Using in-context learning allows a user to “fine-tune” the LLM for the specific tasks being performed in the application [204].

2) Negative Prompting: Negative prompting [205], [206], [207] provides directions to the LLM about aspects of the prompt that it should avoid generating or deliberately excluding during the generation process [205]. 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 [208]. 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. "Don’t write anything that is offensive or harmful, or factually incorrect." This prompt tells the model to avoid generating text that could be offensive or harmful to others and inaccurate. Notably, the authors in [209] conducted experiments for text based image tranlation and found that negative prompting to be very useful when working with textureless images. Moreover, this type of prompting is very useful when working on text to image generation scenarios and has been incorporated in text to image generation methods such as Muse [210].

3) Visual Prompting: Visual prompting [211] 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 to provide the AI model with a starting point or an example/reference that it can use for the given generative task. For images, this may be given to modify the image provided or generate something that is similar in style, color, or texture etc [212]. 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 [213]. Visual prompting provides greater control of the generated output and therefore results in a more accurate result. 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.) [98], the composition of music (wherein the supplied music piece can be used as a reference for the type of music to compose) [214], game development [215] (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) [216].

III. Overview of LLMs

Large Language models have been a key driver for the harbinger of the generative AI revolution.

Language Models have transformed the way we interact with and process language, opening up new possibilities for natural language understanding, generation, and communication [217]. They continue to evolve, pushing the boundaries of what is possible in the realm of language processing and artificial intelligence [218]. In this Section, we briefly discuss the history, evolution, and training of LLMs.

A. History of LLM

The history of LLMs can be traced back to the early days of NLP research [219], [220]. The first language models were developed in the 1950s and 1960s [221]. These models were rule-based [222] and relied on hand-crafted linguistic rules and features to process language [223]. They were limited in their capabilities and were not able to handle the complexity of NLP [224].

In the 1980s and 1990s, statistical language models were developed [31]. These models used probabilistic methods to estimate the likelihood of a sequence of words in a given context [225]. They were able to handle larger amounts of data and were more accurate than rule-based models [226]. However, they still had limitations in their ability to understand the semantics and context of language [227].

The next major breakthrough in language modeling came in the mid-2010s with the development of neural language models [228]. These models used deep learning techniques to learn the patterns and structures of language from large amounts of text [10]. The first neural language model was the recurrent neural network language model (RNNLM) [42], which was developed in 2010. RNNLM was able to model the context of words and produce more natural-sounding text than previous models [229]. In 2015, Google introduced the first large-scale neural language model called the Google Neural Machine Translation (GNMT) system [230].

The development of LLMs continued with the introduction of the Transformer model in 2017 [45]. The Transformer was able to learn the longer-term dependencies in language and allowed for parallel training [231] on multiple Graphical Processing Units (GPUs), making it possible to train much larger models [232]. The release of OpenAI's GPT-1 [233] 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 [234]. While 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. 4).

In 2020, OpenAI released GPT-3 [235], which was able to generate highly coherent and natural-sounding text [236]. GPT-3 demonstrated the potential of LLMs for a wide range of NLP tasks [237]. Inspired by the success of GPT-3, OpenAI released the next iteration of their language model, GPT-4 [238] with the ability to generate even more coherent and natural-sounding text. Following GPT-4's success, Meta also introduced Llama [15], a family of open-source foundation models. Google introduced Bard [239], Amazon introduced AI features in the Alexa [240] models, and Huawei introduced Pangu models [144], joining the AI race.

B. Training of LLMs

Training large language models involves several key steps that are fundamental to their successful development [242]. The process typically begins with the collection and preprocessing of a large amount of text data from diverse sources [243], such as books, articles, websites, and other textual corpora (see Table. III). The curated dataset [244] serves as the foundation for training the LLMs. After the removal of duplicates [245], noisy and poisonous data [246] and ensuring privacy reduction [247], the training process involves unsupervised learning, where the model learns to predict the next word in a sequence given the preceding context assuming the language generation as a random process [248].

Currently, LLMs utilize Transformers which enable them to model long-range dependencies [249], understand text data [250] enable them to generate new content in the style and characteristics of a genre or author [217]. The training objective is to optimize the model's parameters to maximize the likelihood of generating the correct next word in a given context [85]. This optimization is typically achieved through an algorithm called stochastic gradient descent (SGD) [251] or its variants, combined with backpropagation [252], which computes gradients to update the model's parameters iteratively. Some of the popular transformer-based LLMs are discussed below.

- Bidirectional Encoder Representations from Transformer (BERT): BERT [48] is a prominent language model with significantly advanced NLP tasks. Its training process comprises pretraining and fine-tuning stages [253]. During pretraining, BERT learns a general language representation from large-scale unlabeled text data. It employs masked language modeling (MLM) [254] and next-sentence prediction (NSP) tasks [255].
  MLM involves masking a portion of input tokens and training the model to predict the original masked tokens, fostering bidirectional context understanding [256]. 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 [257]. It employs backpropagation and gradient descent optimization to update model parameters.
  Training BERT demands significant computational resources [234], utilizing high-performance hardware like GPUs or Tensor Processing Units (TPUs) or field programmable gate arrays (FPGAs) [258], [259], [260]. Techniques such as layer normalization [261], residual connections [262], and attention mechanisms inherent in the transformer architecture further enhance BERT's

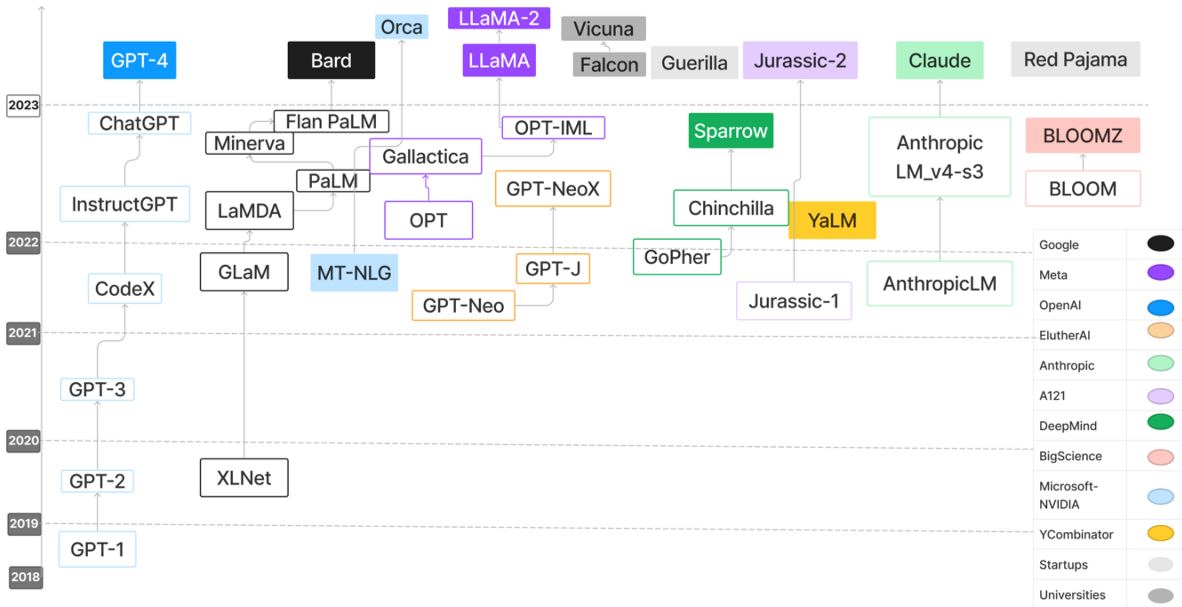


Fig. 4: Illustration of the evolution of Large Language Models (LLMs) over time, highlighting their development across a range of research and commercial organizations. Starting from the initial advancements made in this field, the figure maps out the journey of LLMs, outlining the key milestones, breakthroughs, and model iterations along the way.

TABLE III: State-of-the-art for LLM training pipeline [241]. 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

capacity to capture intricate dependencies and long-range contextual relationships.

- eXtreme Language understanding Network (XLNet): XLNet [263] is a generalized autoregressive [264] pretraining 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 [265]. 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 [263].
  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 [266]. Unlike traditional autoregressive models that can rely on the causal order of words for prediction [267], 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 [85], [268], [269]. 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 [85].
- Text-to-Text Transfer Transformer (T5): T5, developed by Google, is a versatile language model that is trained in a "text-to-text" framework [270]. 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 [271]. 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 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 [270].

- Conditional Transformer Language Model (CTRL): CTRL is a language model designed to generate text based on specific control codes or prompts [272]. 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 [273].

IV. Taxonomy of Large Language Model Tasks

LLMs have a wide array of uses for the tasks of processing natural language including but not limited to writing, summarization, translation, retrieving information as shown in Fig. 5. In this section, the various tasks of LLMs towards working with developing such systems have been discussed.

A. Question-answering

Question-answering (QA) systems [274] allow users to obtain direct answers to questions posed in natural language. LLMs have become a key component in building robust QA systems [275]. LLMs can be effectively pretrained on large text corpora and then fine-tuned on QA labeled datasets [276]. This adapts them to extract or generate answers from passages of text. The broad linguistic knowledge learned during pretraining allows LLMs to understand the semantics of questions and use that to reason about potential answers. Fine-tuning on QA data teaches the models to identify relevant context passages and output the correct response [274]. Key benefits of using LLMs include handling complex questions, synthesizing answers from multiple context documents, and generating clarifying responses when a query is ambiguous. LLM-based QA systems have achieved high accuracy on benchmark datasets [277], surpassing previous state-of-the-art methods. They can be deployed via voice assistants [278], search engines [279], and other interfaces to provide users with quick access to information through natural dialog. Ongoing research is focused on improving reasoning abilities, explainability, and efficiency of LLM question answering.

B. Text Generation

Text generation is a useful application of large language models, which can automate the process of generating content for various purposes [280], such as articles [281], blogs [282], research papers, social media posts, product descriptions, source codes, emails, and more. With their ability to comprehend and generate natural language, these models can produce high-quality content that is both accurate and coherent [283].

C. Language Translation

LLMs possess the capability to translate text from one language to another with exceptional accuracy and fluency [284]. This feature is beneficial for a range of users, including language service providers, global companies, and individuals, who can utilize these models for real-time translation, localization, and overcoming language barriers in communication [285]. The impressive accuracy and fluency of these models make them a valuable tool for facilitating effective communication across different languages and cultures [286]. This feature has the potential to enhance global collaboration and increase access to information, making it an important area of research and development in the field of NLP [287].

D. Text Classification

In addition to their text generation and translation abilities, LLMs are also equipped with exceptional organizational capabilities, such as text classification, analysis, and categorization based on predefined labels or topics [288], [289]. This feature enables the models to effectively manage large volumes of textual data, making them highly valuable for a range of tasks such as sentiment analysis [290], spam detection [291], content moderation [292], and customer feedback analysis [293]. By automating these processes, language models can streamline data management [294], reduce manual labor, and improve the accuracy and efficiency of analysis [295]. These capabilities are particularly useful for businesses and organizations that deal with large amounts of textual data and require effective methods for organizing and analyzing it.

E. Summarization

LLMs possess the ability to generate concise and coherent summaries of lengthy texts or documents [296]. This feature is highly advantageous for a variety of uses, such as summarizing news articles, research papers, legal documents, and other types of content where extracting essential information is crucial. Summarization by language models can save time and effort while ensuring that the most important points are captured accurately [297]. This feature has the potential to enhance the efficiency and effectiveness of content absorption and subsequent creation, making it a valuable tool for individuals and organizations.

F. Virtual Assistance

In the realm of virtual assistants and chatbots [298], LLMs play a critical role. These models possess the ability to comprehend user queries, provide relevant information, and engage in natural language conversations. This capability enables virtual assistants and chatbots to assist with customer support, offer personalized recommendations, answer questions, and automate routine tasks, thus enhancing user experiences and increasing operational efficiency. By leveraging large language

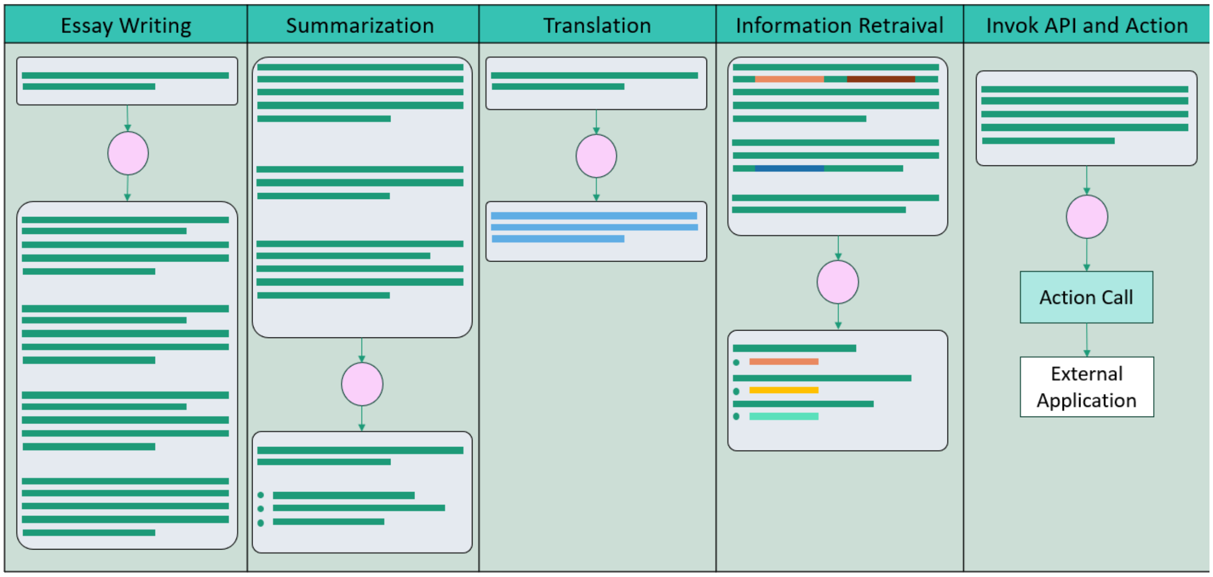


Fig. 5: Examples of LLM. Prompts can be to write an eassay on X topic, summarize a paragrah, translate into X language etc.

models, virtual assistants and chatbots can provide highly effective and responsive support to users while also reducing the workload for human operators [299]. This area of research and development is of significant importance, as it has the potential to transform the way users interact with technology and improve the effectiveness and efficiency of customer support and service delivery.

G. Information Extraction (IE)

The use of LLMs in IE is significant for populating knowledge bases. By leveraging fine-tuned LLMs, entities such as people, organizations, and locations, as well as the relationships between them, can be accurately extracted from unstructured text. This process can facilitate the creation of structured knowledge graphs that can be utilized for various purposes [300]. In addition, LLMs assist in event extraction, enabling the identification of key occurrences described in text documents [301]. This feature has the potential to enhance the efficiency and accuracy of information extraction, making it a valuable tool for businesses and organizations that deal with large amounts of textual data. Further research and development in this area can lead to improvements in the quality and effectiveness of IE using LLMs.

H. Dialog systems

In the context of dialog systems, large language models play a crucial role in facilitating language understanding. The development of large pretrained models like Google's Meena and Microsoft's Blender has led to significant improvements in the naturalness and coherence of open-domain chatbots [302]. These models possess the ability to generate informative, interesting, and harmless responses, making conversational agents much more usable. The application of LLMs in dialog systems has the potential to transform the way users interact with technology, creating more engaging and effective conversational experiences [303]. Further research in this area can lead to improvements in the quality and effectiveness of dialog systems, making them even more valuable for a range of applications and industries.

I. Semantic Search

In the field of Semantic Search, query understanding is of utmost importance, and LLMs are unparalleled in their ability to discern the underlying intent and meaning of user search queries [304]. This ability enables next-generation search capabilities that go beyond simple keyword matching. For instance, LLMs can recognize that the phrases "best budget laptop"; and "affordable student computer" convey the same information need. This feature has the potential to enhance the accuracy and relevance of search results, making it easier for users to find the information they need. Further research and development in the area of Semantic Search can lead to the creation of more effective and efficient search systems, making LLMs a valuable tool for a range of applications and industries [57].

J. Speech recognition

Automated speech recognition is a crucial aspect of voice interfaces and transcription. While traditional systems relied on hidden Markov models or Gaussian mixture models, the emergence of deep learning has seen large neural network models like LLMs take center stage in state-of-the-art results. LLMs that are pretrained on massive text corpora offer rich linguistic knowledge pertaining to language structure, context, and word relationships. Fine-tuning these models on transcribed speech data using connectionist temporal classification loss enables them to learn acoustic-to-text mappings. This leads to significant improvements in the accuracy of automated speech transcription, even in the presence of accented speech or domain specific vocabulary [305]. The contextual knowledge and continual learning abilities of LLMs make

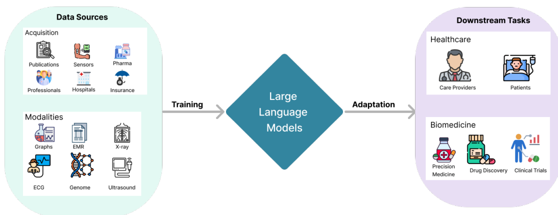


Fig. 6: Illustration of an interactive framework where LLMs enable various tasks across healthcare and biomedicine when trained on multimodal data generated by various sources in the healthcare ecosystem.

them ideally suited for handling the variability and ambiguity inherent in speech signals. As LLMs continue to increase in scale, they are becoming the standard for building high performance and robust automated speech recognition systems [306].

To sum up the discussion on different LLMs, Table IV provides information on the performance of various LLMs on different reasoning tasks.

V. Applications of Large Language Models

Given LLMs wide range of applications, we provide a discussion of their use in the fields of medicine, education, finance, media, law and engineering. The selection of these fields 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 [260], [315]. An interactive framework for the integration of LLMs in the healthcare ecosystem is shown in Figure 6.

A. Medical

LLMs like ChatGPT have exhibited remarkable potential in diverse healthcare applications [316]. They have been successfully employed in medical education, radiologic decision-making, clinical genetics, and patient care [317],[318]. In medical education, ChatGPT has emerged as an interactive tool that aids learning and problem-solving [319], [320]. Notably, ChatGPT's performance in the United States Medical Licensing Exam (USMLE) was comparable to or exceeded the passing threshold, without requiring specialized training or reinforcement [319]. Moreover, ChatGPT's explanations displayed a high level of concordance and insightful understanding [317].

In [277], introduced MultiMedQA, a new benchmark dataset for evaluating LLMs on clinical tasks. MultiMedQA combines six existing medical question-answering datasets spanning professional medicine, research, and consumer queries. They introduced the concept of instruction prompt tuning [321], which can be used to improve the performance of LLMs on a variety of clinical tasks.

In [322], the potential of ChatGPT in radiologic decision-making is emphasized, showcasing its feasibility and potential benefits in enhancing clinical workflow and promoting responsible utilization of radiology services. Similarly, Kung et al. [317] 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, Duong and Solomon [323] 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 [324] evaluated 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 neurosurgical research and patient care, LLMs have been investigated for their potential role in various aspects, including gathering patient data, administering surveys or questionnaires, and providing information about care and treatment [325]. 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. Furthermore, AI-powered chatbots 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 [326].

There are several tools already in use that allow the system to interact with patients such as Ada Health [327], Babylon Health [328], and Buoy Health [329]. 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 [330], 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. Another big development is the segment anything (SAM) model by meta, which may be fine-tuned for a variety of medical image tasks [331]. In the drug discovery domain, DrugGPT [332] is developed, which can be used to design potential ligands, targeting specific proteins, using text prompts.

1) Foundation models for generalist medical AI: The authors in [333] propose a new paradigm for medical AI called generalist medical AI (GMAI). GMAI models are trained on large, diverse datasets of medical data, and they are able to

TABLE IV: Comparison of LLMs' Reasoning Performance. Notations: MMLU [307]: high school and college knowledge, GSM8K: elementary school math, MATH: very hard math and natural science. All current models struggle, BBH [308]: a collection of 27 hard reasoning problems, HumanEval [309]: a classical dataset for evaluating coding capability, C-Eval [310]: a collection of 52 disciplines of knowledge test in Chinese, TheoremQA [311]: a question-answering dataset driven by STEM theorems. [312], [307], [15], [313], [314], [310]

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	-	-	-

perform a wide range of tasks, such as diagnosis, prognosis, and treatment planning. GMAI models are able to outperform traditional medical AI models on a number of tasks, including diagnosis, prognosis, and treatment planning. They can produce expressive outputs such as free-text explanations, spoken recommendations or image annotations that demonstrate advanced medical reasoning abilities and help medical professionals.

B. Education

The impact of AI on education has been a topic of much discussion in recent years since it has emerged as a promising tool for education, with applications including providing meaningful feedback to students [334], aiding teacher improvement [335], and designing personalized and adaptive learning experiences tailored to individual students' needs [336]. However, implementing tech solutions to effectively scale quality education inclusively is an immense task. In this context, general-purpose Large models that are applicable across multiple tasks and subject areas offer potential solution [337].

Foundation models have already begun improving performance in certain educational tasks. For instance, MathBERT [338] has been utilized for "knowledge tracing" – tracking a student's understanding over time based on past responses, and for the "feedback challenge" – interpreting a student's answer to a structured open-ended task. The question is, can foundation models trigger even more significant changes in education? And what are the associated risks of applying these models in an educational context? It's crucial to initiate the discussion around these models with ethical considerations [339].

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 [340] [341] [342]. The accuracy rate for the exams discussed in [340] was below 70 percent indicating its inability to pass the AHA exams. However, this conclusion was drawn due to a design limitation in their study, where they only generated a single response using ChatGPT, introducing bias and severely underestimating ChatGPT's capabilities in this domain. However, the latest study revealed that ChatGPT's accuracy rate increased to 96 and 92.1 percent for the Basic Life Support (BLS) and Advanced Cardiovascular Life Support (ACLS) exams, respectively, allowing ChatGPT to pass both exams with outstanding results [343].

One of the main advantages of using ChatGPT and AI bots in education is that they can help students complete their assignments more efficiently [344]. 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 [345].

This can help students 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 [346]. The goal of incorporating ChatGPT into their platform is to enhance tutoring and coaching experiences by providing one-on-one interactions with students. Khan Academy [4] has been an early adapter of GPT-4 based LLMs working as online tutors, becoming the largest case study for the evaluations of LLMs in an educational context in the process.The incorporation of AI in tutoring and teaching proves that it can be a valuable tool in reducing negativity, particularly the perception that its main purpose is for cheating. Undoubtedly, AI technology is still in its nascent phase, yet it shows great potential in supporting students and catering to their individual requirements. [347].

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 [348]. 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 [344].

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 [349]. 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 the 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. It is crucial to address the issue of inadequate training and contextual fine-tuning for LLMs, as their potential utilization without such preparations raises significant concerns [350]. While it is true that LLMs possess the ability to provide answers to a wide range of questions and assist users in generating responses effortlessly, it is essential for students and scientists [351] to receive adequate training specific to their needs in order to fully harness the capabilities of LLMs. Neglecting the necessity for context-specific training and fine-tuning can render these tools less effective and limit their true potential.

Another concern is that the use of AI bots in education could lead to increased inequality [352]. 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 [353]. In conclusion, the use 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.

[4]https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/

C. Finance

LLMs are making significant advancements in the finance industry [354] with applications ranging from financial NLP tasks [355], risk assessment, algorithmic trading [356], market prediction [357] and financial reporting [358]. LLM's such as BloombergGPT[99], a 50 billion parameter large language model trained on large diversified financial corpus, has revolutionized financial NLP tasks including but not limited to 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 [359], loan approvals and investments [360]. Algorithmic Trading [361] is another application that can leverage LLMs 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 [362] which is an open-source LLM tailored for finance. It is expected that more work will be carried out in this domain.

D. Engineering and similar

LLMs have gained substantial attention across various fields, and their potential applications in engineering domains are increasingly being explored [363]. For instance, ChatGPT has diverse applications in software engineering, including code generation, debugging, software testing, NLP, documentation generation, and collaboration.

In software engineering, CodeGPT 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 [364]. Additionally, CodeGPT 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 possibility of ChatGPT utilization to various calculations in mechanical engineering was attempted in [365]. However [365] 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 this study, 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 potentially have severe negative consequences [366].

In manufacturing, Wang et al. [367] conducted an evaluation of ChatGPT's capabilities in supporting design, manufacturing, and engineering 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. However, it was found that, in manufacturing, ChatGPT struggles to understand questions sometimes 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. [368], 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. 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.

E. Media and Entertainment Industry

The media and entertainment sector is currently undergoing a transformative phase that revolves around data and prioritizes consumer-centric experiences [369]. Companies of all sizes are now striving to introduce groundbreaking innovations that enable personalized, one-to-one interactions on a large scale [370], [371]. Moreover, content lies at the core of the Media and Entertainment industry, and its creation is evolving with the integration of data, particularly social signals, into content strategies [372]. LLMs have a transformative role, revolutionizing how companies leverage data and AI for content development and curation. LLMs assist in generating captivating headlines, compelling copy, and providing real-time content feedback, streamlining production and enhancing quality. They have come out as a game changer among the various technologies driving the technological revolution in the media and entertainment industry. LLMs not only enable the creation of original content but also demonstrate a profound grasp of intricate information and the ability to simulate human-like interactions. This includes MediaGPT, a large language model for the Chinese media domain which was presented recently. It can generate high-quality and relevant outputs for various tasks in the Chinese media domain [373]. Similarly, Robertuito [374] was proposed for Spanish social media.

Large AI models can also be utilized for generating attractive advertisements and marketing [100], political speeches, slogans and social media posts [375], and promotional videos [117]. Similarly, leading entertainment networks and applications are using LLM based algorithms that can analyze user data to offer personalized recommendations for movies, TV shows, and music. This helps entertainment companies to retain customers and improve their engagement with their content [376]. Moreover, LLMs automate content curation fostering user satisfaction, retention, and monetization. Recently, many companies have developed and offered their services for media and entertainment purposes. One of the prime examples of such services is Dolly, an LLM-trained model developed by databricks Incorporation [377].

The creation of AI-based newscasters [378] is a recent concept that consists of virtual news presenters or anchors that are generated using AI technologies, particularly LLMs [379]. In April 2023, a Kuwaiti media outlet unveiled a virtual news presenter "Fedha" with plans for it to read online bulletins [380]. At the University of Kent's Centre for Journalism, lecturers are grappling with how to prepare the next generation of reporters for the potentially AI-powered newsrooms of the future [381]. AI algorithms have the capability to analyze user data, providing tailored suggestions for movies, TV shows, and music. This enhances customer retention and boosts engagement with entertainment content. Table V presents the recent tools and applications that are transforming the entertainment industry.

F. Role of LLMs in the Future of Legal Practice

With advancements in AI and the development of tools such as GPT-4, Bard, and Bing, it is aimed that these advancements will empower lawyers to enhance legal research, drafting tasks, and decision-making [382]. This has sparked interest among entrepreneurs developing AI tools [383], law firms integrating AI into their workflow, and law professors exploring AI-based techniques for legal aid [384]. A recent example is Chatlaw [385] model, which is open-source legal language model. A legal informatics approach was introduced in [386] to align AI with human goals and societal values. By incorporating legal knowledge and reasoning into AI systems, the paper contributes to the research agenda of enhancing the integration of AI and law.

In [387], the authors propose legal prompt engineering (LPE) as a means to improve LLM performance in legal judgment prediction tasks. The effectiveness of this method has been demonstrated on three multilingual datasets, showcasing the model's capability to handle the intricacies of legal language and reasoning from various sources of information. LLMs' transformative potential in the legal field is evident from their impressive performance in legal exams. GPT-4 scored in the 90th percentile on the Uniform Bar Examination

TABLE V: Unveiling the AI Revolution in Entertainment: Real-world Illustrations

Tools	Function	Link
Scriptbook	A cutting-edge AI-powered script analysis tool, is harnessed by film studios to forecast the commercial triumph of a screenplay. The tool meticulously assesses the script's characters, themes, and plot points, drawing comparisons to the historical performance of comparable films to foresee its potential box office success.	Scriptbook
Aiva	AIVA (Artificial Intelligence Virtual Artist) represents an AI-driven music composition tool that generates original music tracks tailored to user preferences. By analyzing data points such as genre, tempo, and mood, the tool crafts unique compositions suitable for integration into films, TV shows, and video games.	Aiva
LyricFind	LyricFind takes center stage as an AI-powered lyrics search engine, empowering users to find song lyrics using natural language queries. By employing natural language processing algorithms, the tool comprehends user queries and delivers precise and relevant results.	LyricFind
Ziva Dynamics	Ziva Dynamics showcases an AI-powered software tool tailored for creating authentic 3D character models in films and video games. The tool utilizes machine learning algorithms to simulate muscle and skin movement, resulting in character models that boast unparalleled realism and intricate detailing.	Ziva
DeepMotion	DeepMotion introduces an AI-powered animation tool capable of producing lifelike 3D animations for video games and films. Leveraging machine learning algorithms, the tool simulates human movement and behavior, delivering animations with enhanced realism and natural aesthetics.	DeepMotion
Speechify	Speechify is one of the most popular and efficient first AI Voice Over generators for using famous singer's voices for singing differents songs. It also creates human-quality voice-over recordings in real time. Narrate text, videos, explainers anything you have and in any style.	Speechify

[61], and ChatGPT autonomously passed four law school final exams at a top law school [388]. These achievements showcase the significant impact of AI language models on legal practice. The authors present Chain-of-Thought (CoT) prompts, which aid LLMs in generating coherent and contextually relevant sentences following a logical structure, simulating a lawyer's analytical approach [389]. The study shows that CoT prompts outperform baseline prompts in the COLIEE entailment task using Japanese Civil Code articles. LLMs have also been utilized to explore fiduciary obligations [55].

In a recent working paper by Choi et al., the authors conducted experiments using ChatGPT to generate answers for four authentic exams administered at the University of Minnesota Law School [390]. In summary, the authors concluded that ChatGPT successfully passed all four exams with an overall average grade of C+. This level of performance would grant it credit towards a JD degree, but it would also place the student on academic probation. Interestingly, if ChatGPT maintained this performance throughout law school, it would be able to graduate successfully. However, ChatGPT's answers exhibited consistent issues and errors, which rendered its performance significantly poorer compared to the average student. One of its main challenges was "identifying and addressing issues" when presented with open-ended prompts, a crucial skill in law school exams.

Recently in June 2023, in response to fake case citations generated by ChatGPT and submitted in a court filing, a US judge has imposed a fine of $5,000 (3,935) on two lawyers, along with their law firm Levidow & Oberman [391] [392], due to the fake citations generated by ChatGPT. The fictitious legal research was utilized in an aviation injury claim, and the lawyer admitted to inventing six non-existent cases citations. In Texas, a judge now requires attorneys to verify that no part of a filing was composed by generative AI or, if it was, that a human has verified its accuracy [393] [392]. However, not all judges share the same stance on chatbots in legal proceedings. For instance, Judge Juan Manuel Padilla, based in Colombia, acknowledged using ChatGPT's assistance in a case concerning an autistic child [393].

In light of these examples and use cases, LLMs indeed offer numerous benefits, but it is crucial to recognize and comprehend their limitations. They can serve as a valuable tool for initial research, explanations, and improving efficiency in legal practice. However, lawyers must be mindful of its limitations. While they can provide seemingly convincing answers, they may still be misleading or inaccurate. Their reliance on statistical patterns from training data means they lack human-like reasoning and may not incorporate the most recent legal developments. Ethical and confidentiality concerns also arise due to the storage and potential use of prompts and information for training purposes, posing risks to sensitive information. While lawyers will likely need to integrate AI to remain competitive, it must be done responsibly, upholding ethical obligations.

G. Marketing

Large language models are crucial in modern marketing, transforming customer engagement and content delivery [394]. They excel in content generation, creating compelling product descriptions, ad copy, blogs, and social media posts, saving time and resonating with audiences [395] [396]. Personalization is a standout feature, allowing marketers to deliver tailored messages based on customer data, improving satisfaction and loyalty. Customer support is revolutionized by chatbots, providing instant assistance 24/7, and reducing the workload on support teams [397]. In market research, large language models analyze vast data, including feedback and social media, offering insights into trends, sentiment, and competition. They contribute to SEO optimization, identify keywords, and enhance social media monitoring [396].

The adoption of LLMs and ChatGPT in marketing offers numerous benefits, but it also comes with potential risks for marketers, consumers, and other stakeholders [396]. The similarity and lack of uniqueness in ChatGPT's responses to similar prompts from different marketers could undermine the distinct identity of the marketer or brand. This presents a challenge for marketers who prioritize creativity and innovation in their strategic decision-making, especially within an AI-driven environment [398] [395] [399]. AI marketing tools like ChatGPT may draw information from unreliable sources, leading to the provision of incorrect information. In the worst-case scenario, if inaccurate data overwhelms the system, it can lead to false outcomes [396] [400]. Ethics is a significant concern as LLMs can generate content that appears human-generated, raising transparency and disclosure issues [401].

Negative consumer perceptions could arise if AI-generated content is overwhelming or perceived as inauthentic, leading to reduced trust in the brand. Compliance with regulations and data protection laws is crucial to avoid potential legal consequences [402]. Moreover, marketers may become dependent on third-party AI providers, leading to vendor lock-in or reliance on external platforms. To mitigate these risks, marketers must use AI responsibly, provide clear disclosure when AI is involved, and maintain human oversight to verify the accuracy and appropriateness of AI-generated content [403]. Regular audits, continuous monitoring, risk assessment with its mitigation, and adherence to AI ethics guidelines can help ensure that marketers' use of LLMs and ChatGPT aligns with best practices and meets consumer expectations. It is at the interjunction of the risks (market and consumers both) and ethics where we can balance the scale which opens the window of opportunity for responsible AI adoption (see Figure 7).

1) Customer Service : Another application that has witnessed a significant impact is customer service. LLM-powered chatbots and virtual assistants are increasingly being integrated into customer support systems, providing companies with a scalable and efficient means of addressing customer inquiries and concerns [404] [405]. Unlike human representatives, LLM-powered chatbots can process and respond to inquiries instantaneously, enhancing the overall customer experience [406]. The use of LLMs also leads to cost savings for companies [407]. Due to the high cost associated with employing dedicated customer service agents, an increasing number of companies are exploring the use of NLP to assist human agents [408]. NLP enables the auto-generation of responses that can be directly utilized or modified by agents. In this context, LLMs emerge as a natural and suitable solution.

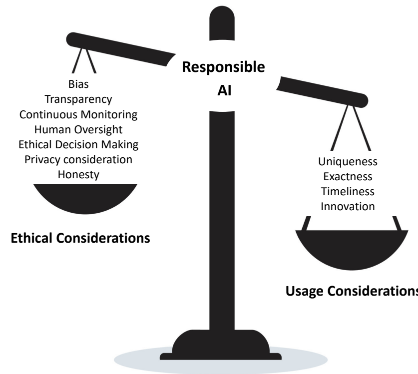


Fig. 7: The Tradeoff for Responsible AI: Example of marketing.

In [406], the authors have introduced a cost framework to evaluate NLP model utility for customer service. They compare three LLM strategies - prompt engineering, fine-tuning, and knowledge distillation - using agent feedback. The system was able to generate dynamic and context-specific suggestions to assist the customer. Similarly, ChatGPT can be trained and fine-tuned with customer-specific data, enabling it to deliver personalized and customized answers for individual customers.

Furthermore, LLMs and ChatGPT/GPT-4 have exhibited exceptional potential in customer service, but they also come with inherent limitations and challenges. Context understanding remains a critical issue, as LLMs and ChatGPT/GPT-4 may struggle to grasp complex queries fully, resulting in responses that lack nuance and accuracy. Additionally, emotional intelligence, scalability, and the preference for human interactions represent further challenges in delivering seamless customer service experiences [409].

VI. AI-Enabled Tools: The Future of Everything

AI tools are becoming increasingly powerful and versatile. They can be used to generate text [410], images [411], and videos [412], translate languages [286], write different kinds of creative content [413], and answer the users questions in an informative way [414]. These powerful tools provide a user-friendly interface for the optimization of daily routine tasks [85]. One such example is the popular website, "There's an AI for THAT" [5], which contains about $7\tilde{K}$ AI tools for $2\tilde{K}$

[5] https://theresanaiforthat.com/

different tasks. In this Section, we discuss various AI-enabled tools based on LLMs or text prompts.

A. Chatbots / ChatGPT

Chatbots are frequently used in customer service applications where they can respond to queries, offer assistance, and fix problems [50]. High-tech companies are likely to become even more interested in using chatbots to improve their customer experience and grow their businesses. For example; OpenAI developed ChatGPT [415], Google developed Bard [416], and Meta launched Llama-2 [15]. Here, we critically compare these Chatbots in terms of accuracy, ease of use, cost, integration and others.

1) Comparison between Chatbots: ChatGPT and Google Bard are two of the most popular LLMs available today [417], [418]. The third popular LLM being Bing, which is based on GPT-4. Bard is based on the LaMDA [419] (Language Model for Dialogue Applications) architecture, while ChatGPT is based on the GPT-3 (Generative Pre-trained Transformer 3) architecture. ChatGPT was modified and improved using both supervised and reinforcement learning methods [420], with the assistance of human trainers (RLHF) [80]. The learning includes three steps;(i) supervised fine-tuning [421], reward model [422], and maximum policy optimization [423]. First, a pre-trained GPT-3 model is used and fine-tuned with the help of labelers by creating a supervised dataset. After the supervised fine-tuning, different input prompts are fed to the model, and 4 to 7 responses are generated for each response, the labelers rank each response of the model. The responses are scalar values, which are used to train the reward model. In the third step, the model is tested on unseen input sequences, responses are evaluated by the reward model, and the output reward is used to fine-tune the parameters of the model, to incorporate more human-like characteristics and behaviors via reinforcement learning [424].

LaMDA is a newer architecture (conversational neural language models) that is specifically designed for dialogue applications. Bard uses LaMDA, which is a hybrid architecture that combines batch processing [425] and streaming processing [136]. This allows BARD to handle both historical and real-time data. It is trained on a massive dataset of text and code, while ChatGPT is trained on a massive dataset of text, which means that Bard has a broader understanding of the world and can generate more comprehensive and informative responses, while ChatGPT is better at generating creative and interesting responses.

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. Bard is more personalized than ChatGPT, the responses generated by Bard are more tailored to specific needs, and it is also more scalable than ChatGPT. A comparison between ChatGPT, Bard, and BingChat is made in [416] on VNHSGE [426] dataset, which is a Vietnamese High School Graduation Examination Dataset for Large Language Models.

The results indicate that BingChat performed better than Bard, and ChatGPT. All models perform better than the Vietnamese students [427]. In fact, a comparison between the three popular LLMs services, namely, ChatGPT, Bard, and Microsoft Bing has been of interest to researchers and field practitioners. A recent work by Bhardwaz et. al [428] provided a general comparison of these three models considering accuracy, response time, user experience, and engagement. From their experiments, they found that ChatGPT provided the most relevant responses and accuracy, Bard provided the quickest response and Bing provided the best user experience and engagement. Another comparison by Campello et. al [429] experimented with four different chatbots (above three and Quoras Poe [429]) when asked to solve an intelligence test for recruitment in Brazil found that all four chatbots scored above the 95th percentile while ChatGPT and Bing scored 99th percentile. These are in addition to comparisons being made for typical as well as atypical specific use cases such as news fact-checking (GPT-4 performing the best) [430], taxes [431] as well as political leaning [432] and more. To complete the discussion, Table VIII presents a comparison between ChatGPT, Google Bard, Llama-2, and Microsoft Bing Chatbots.

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

1) Diffusion Models: Diffusion models are a scheme of generative models that have provided excellent performance in a variety of applications, most notably the synthesis of images [433]. Starting from a sample of a target data distribution, a diffusion model works in two steps, a forward diffusion process and a reverse diffusion process. The forward diffusion process gradually adds increasing amounts of Gaussian noise[434] to the sample image successively over time, until it becomes the Gaussian distribution [435]. The model is then tasked to start from this noisy image version and undo the noise addition by going through a reverse process to recreate the original data [436], [20].

The forward process takes the form of a Markov chain [437] where the distribution at a given time instant only depends on the sample from the timestep immediately preceding it. Therefore, the distribution of the corrupted samples at any given point with respect to the original sample is the product of the successive single-step conditionals up till that point.

Moreover, typically the number of passes of noise addition is in the order of a thousand with the increments each time being quite small. This is necessary to ensure that the reverse process of “recovering” the original sample is more achievable as it has been shown that with infinitely small step sizes, the reverse form will be able to achieve the same functional form as the forward process [438]. Diffusion models use this observation in the forward process.

Similar to the forward process, the reverse run is also set up as a Markov chain. The model starts from a Gaussian noise sample and goes through the sample one timestep at a time to remove noise at each step. In essence, the forward process is designed to push the sample out of the original data distribution while the reverse process is designed to learn to bring it back into the original data distribution.

A diffusion model can be interpreted as a latent variable generative model similar to a variational autoencoder (VAE). The forward process can be thought of as producing latent from data and the reverse process is as converting latent to data. However, as opposed to VAEs, the forward process for diffusion models is typically fixed so that only a single network needs to be trained that deals with the reverse process. The objective function is the variational lower bound on the log-likelihood of the data. It consists of a log-likelihood term or reconstruction term minus a KL divergence term [439] also called the regularization term [440]. The log-likelihood terms [441] encourage the model to maximize the expected density assigned to the data. The KL divergence term encourages the approximated distribution to the prior distribution on the latent variable. Moreover, diffusion models can also be directed to sample conditionally based on a variable of interest which can be incorporated as an additional input during training. This has been the reason that diffusion models have shown better performance than Generative Adversarial Networks [442] in a variety of image generation tasks including perceptual quality ([443], text to image generation ([444], image inpainting [445] and manipulation of images[446].

2) Image generation: The images contained in this section were generated by a model incorporating the stable diffusion process in to existing diffusion models as suggested in [447] and uses text to generate photorealistic images. This model was released by stability.ai [448] and was demonstrated to be capable of generating images which were previously difficult to generate, such as images of people with accurate facial features as well as objects with abnormal or impossible shapes. This capability of being able to generate a diverse set of images from a text input without requiring a large amount of data for training and its public availability opened up text-to-image generation usage for a host of applications. Table VI 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*. 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 an image of a woman*.

3) Video Generation using text prommpts: Text-to-video generation is a challenging task that involves generating video sequences from textual descriptions or prompts [412]. The video generation process involves text understanding [449], video scene generation [450], temporal structure and transition [451], and video synthesis [452]. One such model is T2V [453], which is a video generation model using text

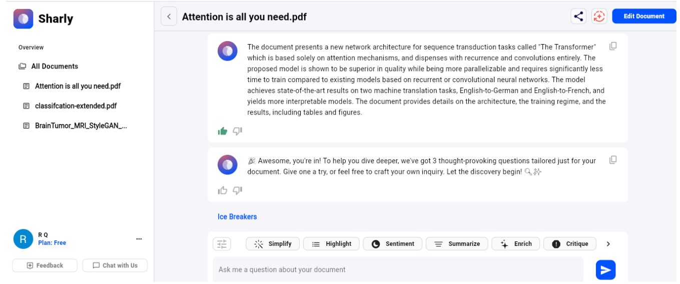


Fig. 8: An example of PdfGPT. Upload any PDF document and start chatting. It helps in summarizing, highlighting, critiquing, and simplifying the content.

prompts.

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 [454]. 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 [455], and Spacy [456]. Similarly, Hugging Face's Transformers library [457] is the state-of-the-art toolkit for developers to implement AI text generation capabilities into their applications; including fine-tune models for sentiment analysis, language translation, and text summarization.. This library offers a collection of pre-trained language models, including GPT-3.5 and various other popular models like Bidirectional Encoder Representations from Transformers BERT [69] and RoBERTa [70].

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 [458], [459]. One such tool is PDFGPT [460], 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. 8.

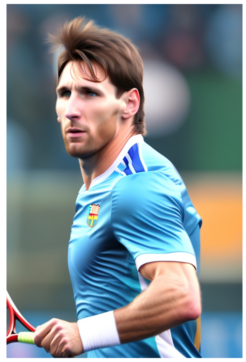
Another interesting AI tool is elicit.org, which helps automate literature reviews. The website offers a variety of features, including, finding relevant literature, summarizing and visualizing literature, and extracting relevant information.

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 [461], [462]. Fake citations is an inherent consequence of the generation capabilities of LLMs wherein they may prefer to lean on their generation capabilities rather than on search. One way to think about this is in terms of sets where the sample space of there being any number of possible citations/references, be them fake or real is large (in terms of word makeup) whereas the pool of real citations/papers which have been published is only a small

TABLE VI: Image generation examples

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


a


b

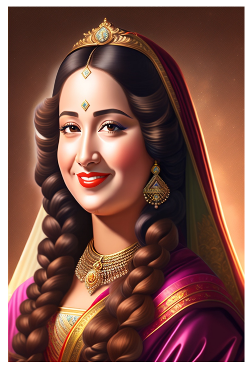
c

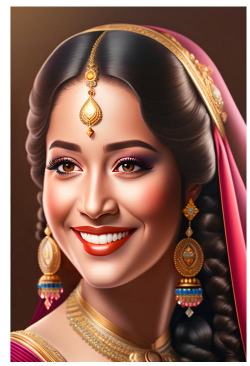

d

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


a

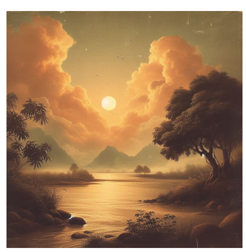

b

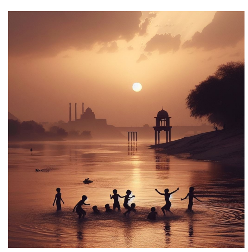

c


d

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


a


b

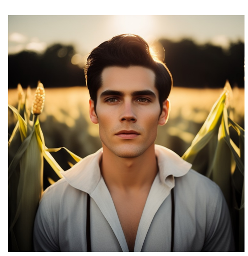

c

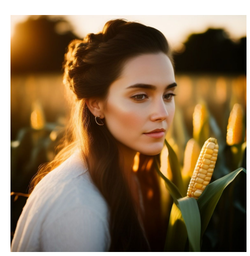

d

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


a

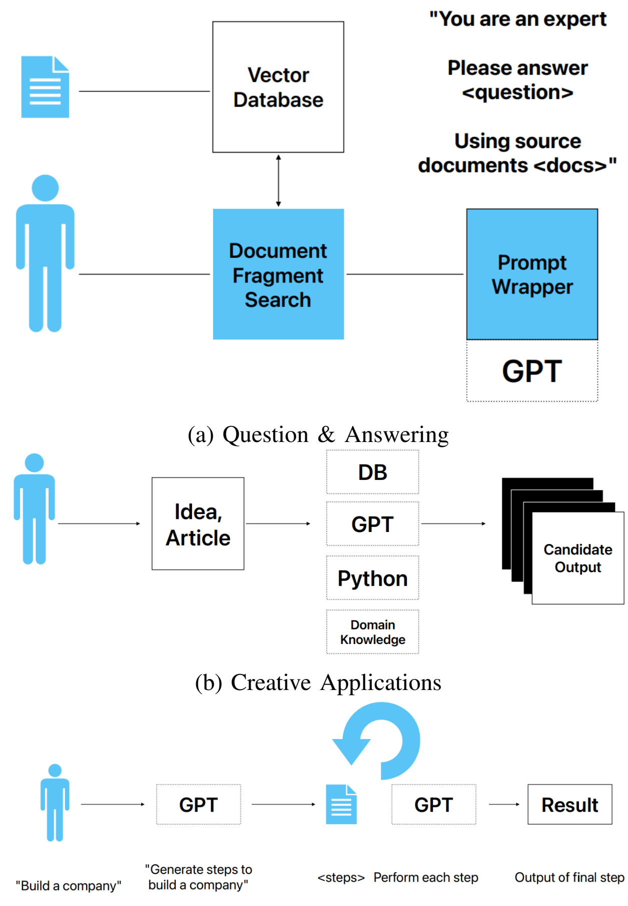

b


c


d

subset of the total possible sample space. So, when an LLM is asked to provide citations, given its inherent generative nature, it might end up generating text that looks like a research article/source but in reality, might just be a sequence of words that are objectively similar in linguistic/language generation terms [463].

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 can undermine the credibility of the author [464]. 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. Finally, developers of AI tools should implement rigorous quality control measures to ensure that their tools generate accurate and reliable citations and references [465].
Recently, WebChatGPT [6] 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.



(a) Question & Answering

(b) Creative Applications

(c) Multi-step Planning

Fig. 9: Templates for LLM-based application development. GPT is taken as an example scenario representing LLMs.

E. AI tools for Coding and Development / 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 [466]. GitHub Copilot [467] 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. Meta also released the Code-Llama [468], a LLM model, that can use text prompts to generate and discuss code. It has the potential to generate clean and robust code with well documentation in Python, c/C++, Java, PHP, Typescript (Javascript), Bash and other programming languages.

Developers can interact with the LLM by providing prompts or queries related to their coding needs, and the model can generate relevant code segments or suggest solutions to programming problems [469]. LLMs have been used to develop applications in three primary categories which include: (a) Question Answering, (b) Creativity (c) Multi-step planning [469]. These template categories are illustrated in Fig. 9. In Table VII, we present a list of publicly available AI/LLM tools for a variety of applications.

VII. 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 [479]. GPT-Plugins can be used to do things, such as access to external data sources, automate tasks, and enhance user experience [480]. In this Section, we demonstrate several GPT-Plug-ins.

A. ChatGPT Plug-ins

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 Chat-GPTs 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 present in its training dataset [422].

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.

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

TABLE VII: Publicly available AI /LLM tools

Tools	Function	Link	Availability
ChatGPT	Conversational AI Chatbot	ChatGPT	Both
RoomGPT	Redesign your room in eight different themes	RoomGPT	Public
HomeGPT	Redesign your home and office	HomeGPT	Subscription based
PDFGPT.IO	Turns PDF into the knowledge base for a ChatGPT type interface	PDFGPT	Subscription based
TexGPT	Harnesses GPT-3's power to help you write in Overleaf	TexGPT	Public
AcademicGPT	An AI tool to write and review scientific papers, critical analysis and explanation of complex concepts	AcademicGPT	Public
DiagramGPT	An AI tool for creating scientific diagrams and flow charts of different processes	DiagramGPT	Public
BloombergGPT	A Large Language Model for Finance	BloombergGPT	NA
AutoGPT	Auto-prompting without the user intervention	AutoGPT	Public
AgentGPT	Autonomous AI agent in the browser	AgentGPT	Public
HuggingGPT [470]	A framework to connect various AI models to solve AI tasks	HuggingGPT	Public
XrayGPT [471]	Automated analysis of chest radiographs	XrayGPT	Public
Video-ChatGPT	A vision language model for video understanding and conservation about videos	Video-ChatGPT	Public
ClimateGPT	Large language model for a conversation about the climate in English and Arabic	ClimateGPT	Public
CodeGPT	An AI assistant for coding	CodeGPT	Public
Code Llama	Open Foundation Models to generate and discuss code	Code Llama	Public
MiniGPT-4 [280]	Multi-modal model for a number of tasks, including image generation and website development, using prompts	MiniGPT	Public
BiomedGPT [472]	A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multi-modal Tasks	BiomedGPT	Public
SkinGPT [473]	An Interactive Dermatology Diagnostic System		
PatientGPT	An AI engine to transform patient navigation by providing healthcare organizations and their patients with a seamless and customized experience	PatientGPT	Subscription based
SentimentGPT [474]	Exploiting GPT for sentiment analysis	SentimentGPT	Public
DrugGPT [332]	A GPT based model to design potential ligands, targeting specific proteins	DrugGPT	Public
Elicit	AI research assistant, automated literature reviews	Elicit	Public
Citation AI	AI research assistant to generate real evidence-based answers	Citation AI	Subscription based
Midjourey AI	AI tool to create realistic synthetic images	Mid Journey	Subscription based
DALL.E2	DALL·E 2 is an AI system that can create realistic images and art from a text description	Daall-e-2	Subscription based
VALL-E	An audio synthesization tool	Vall-e	Public
Gen-2	Video generation using text, images, and videos	Gen-2	Public
AI Avatar	Avatar generation	AI Avatar	Public
Langchain [475]	Building applications with LLMs through composability	Langchain	Public

TABLE VIII: Comparison of Bard, ChatGPT, Llama-2 and Bing Chat

Feature	**ChatGPT** (GPT 3.5)	**Bard**	**Bing Chat** (GPT - 4)	**Llama2**
Accuracy	Not as accurate as Bard	Generally more accurate than ChatGPT	Most accurate	least 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	Less than ChatGPT and Bard both better than Bing
Company	OpenAI	Google	Microsoft	Meta
Primary Purpose	Creative text generation	Conversational AI	Information retrieval	Text generation, answer questions, language translation etc
Integration	Standalone model	Standalone model	Integrated with Bing search engine	Standalone model
Easy to use	User-friendly	User friendly	Not as user-friendly as ChatGPT or Bard	User-friendly
Access to online data	No, trained on data available till 2021	Yes	Yes	Yes
Cost	GPT 3.5 free / GPT-4 (20 USD per month)	Free	Free	Free
Availability	Publicly available	Publicly available	Publicly available	Publicly available
Architecture	Generative pre-trained transformer [476]	Pathways Language models (PaLM2) [477]	Next Generation GPT [478]	Transformer
Plagriasm detector	Yes	No	No	Less likely to generate plagiarised text
Limitations	May generate less coherent or incorrect text	Not as creative as ChatGPT	May provide limited or incomplete information	Trained on a smaller dataset than ChatGPT and Bard, may not generate text for some topics

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 [481]. 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 [7], 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 drawn a big attention towards plugin development. Apart from the above-mentioned early developers, two notable plugins already made available by OpenAI are the Code interpreter and the knowledge-based retrieval plugin.

- **Code Interpreter:** 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.
- **Knowledge-base retrieval:** A knowledge-based retrieval plugin has also been open-sourced[8] which can be used by developers. 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.

Lastly, third-party plugins are also an option. These can be created and have been created by several entities. Fig. 13 demonstrates 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 IX provides a list of plugins available for ChatGPT.

VIII. Guidelines for effective use of Large Language Models

In this section, we provide a list of steps as well as guidelines for effective and responsible use of LLMs [488].

A. Model selection and deployment guidelines

- **Identify the task**: Determine the task, LLMs can be used for a wide range of tasks, such as text classification, sentiment analysis, question answering, and text generation [489], [490], [491].

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

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

TABLE IX: Some ChatGPT Plugins. This list is not exhaustive and more and more plugins are being developed.

Name	Task	Example use cases
Language Translation [286]	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 [482]	Determine tone of text or conversation	This can be used for the task of market research, customer analysis and social media monitoring
Spell Checker [483]	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 [484]	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 [485]	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 [486]	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 [487]	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

- **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 [250].
- **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 [492].
- **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 [493].
- **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 [494].
- **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 [415].
- **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 [495].

B. Ethical guidelines

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 [80].

- **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 [496].
- **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 [497].
- **Address Ethical Implications**: LLMs have the potential to be used for harmful purposes, such as spreading disinformation or generating deepfakes [498]. Ethical considerations; including, ensuring accountability, transparency, and responsibility in the development and deployment of models must be taken into account [499].
- **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 [364].
- **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 [500].
- **Encourage Collaboration**: Collaboration between researchers, developers, and industry should be encouraged to promote the responsible development and use of LLMs. This includes open sourcing models and data, as well as facilitating the sharing of research findings and best practices [501].
- **Minimize Environmental Impact**: Training LLMs can require significant computational resources and energy, which can have negative environmental impacts. De-

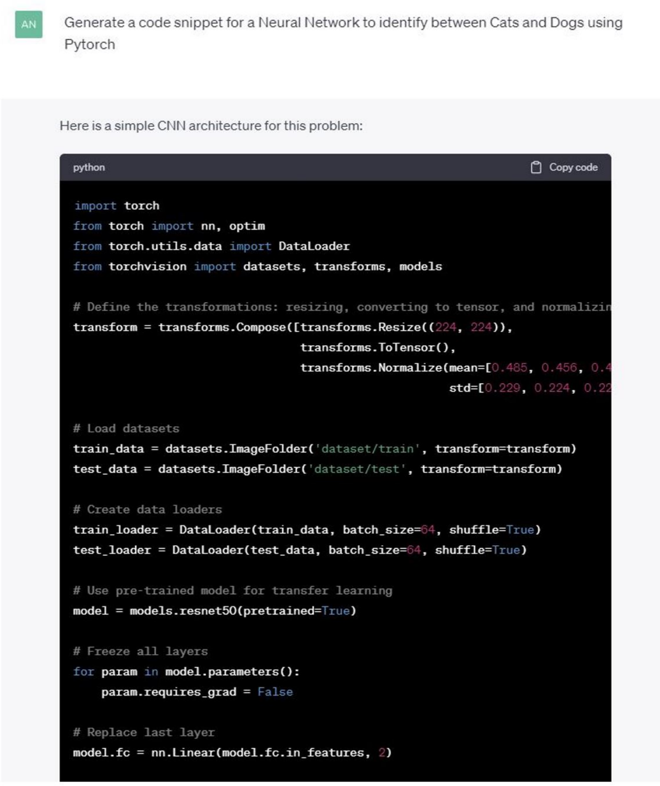


Fig. 10: Demonstration of code generation using ChatGPT

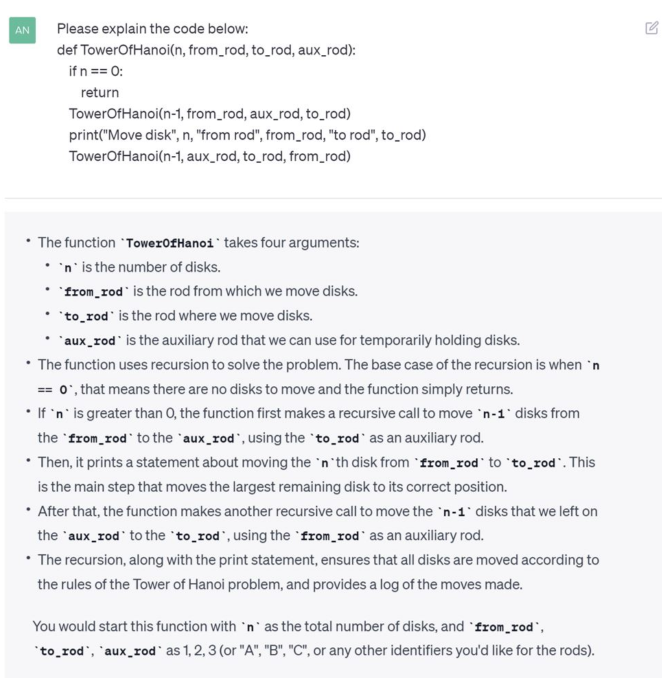


Fig. 11: Demonstration of code debugging using ChatGPT

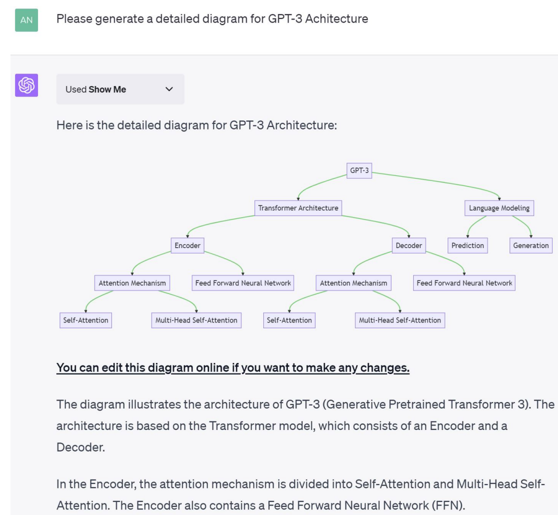


Fig. 12: Demonstration of the ShowMe plugin.

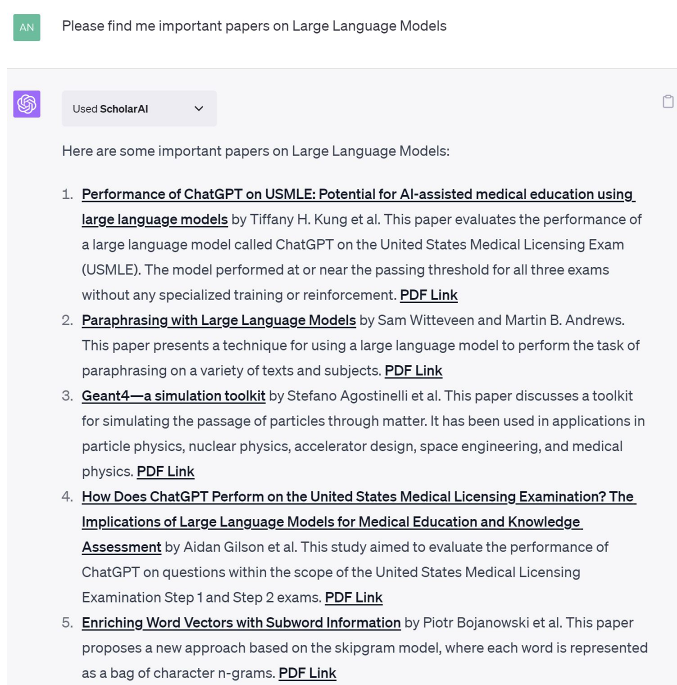


Fig. 13: Demonstration of the ScholarAI plugin.

velopers 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 [502], [503].

- **Optimization is exploitation:** is a statement that holds particular significance in the context of LLMs and AI technologies [504]. 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 [505]. Therefore, it is important to carefully consider the ethical implications of optimization in the development and deployment of LLMs and AI technologies [77].

IX. Challenges and Limitations of Large Language Models

Although LLMs have made significant contributions to various domains, they have significant limitations and challenges [105], [114]. LLMs are currently perceived as forerunners of Artificial General Intelligence (AGI). However, despite their phenomenal success in conversational tasks, the state-of-the-art LLMs still lack in many aspects that makes them less likely an early manifestation of AGI. We first provide a quick list of the challenges and limitations of LLMs (Fig. 14) and then present a more detailed discussion on a few limitations of critical concerns.

A number of challenges and limitations have been focused on, including biased data, overreliance on surface-level patterns, limited common sense, poor ability to reason and interpret feedback [506], [507]. Other issues include; the need for vast amounts of data and computational resources [508], limited generalizability [509], lack of interpretability [510], difficulty with rare or out-of-vocabulary words, limited understanding of syntax and grammar [511], and limited domain-specific knowledge [512].

The susceptibility to adversarial attacks [242], ethical concerns [75], difficulty with context-dependent language [249], absence of emotion and sentiment analysis [513], limited multilingual capabilities [514], limited memory [515], lack of creativity [413], and restricted real-time capabilities [489] are also critical concerns. The 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 [516], [517], [518], [519], [520], [521], [92] are some of the key challenges in safe, responsible, and efficient deployment of LLMs.

A. Training Data Requirements

Large Language Models (LLMs) require a large corpus of data for pre-training the model. Collecting and curating these datasets can be extremely challenging. The size of the dataset makes it impossible to read or assess the quality of the dataset making it prone to having duplicates, making the model biased and degrading its responses [522] [523]. It also makes it difficult to assess the model as the training data may contain data similar to testing samples leading to incorrect evaluation metrics. Since there is no way of checking the datasets manually, it may contain confidential or personal information too, such as telephone numbers leading to privacy leaks during prompting [506]. Due to the fact that the data distribution and requirements in LLMs is more of a black box, it remains uncertain what amount of data is required for different tasks.

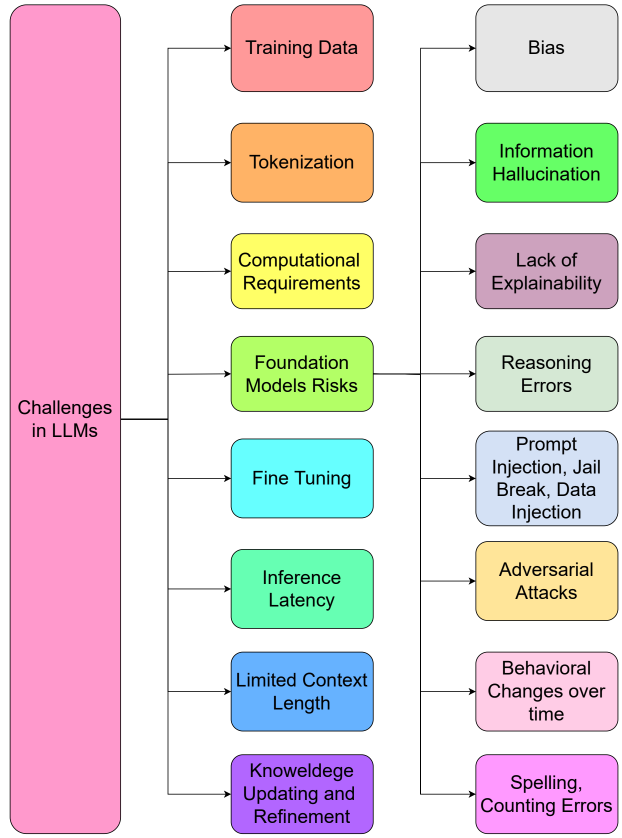


Fig. 14: Challenges in LLMs.

B. Tokenization Problems

LLMs heavily rely on tokenization which consists of breaking down a sequence of words into tokens for the models input. Most LLMs use subward tokenization [56], which is used to create tokens by splitting the words to handle non-familiar vocabulary and at the same time maintaining the computational complexity. However, there are some major drawbacks of tokenization which includes, different combinations of token can be used to relay the same prompts, which may lead to unfair pricing for the APIs of these LLMs. In a multilingual environment it may cause unexpected model responses due to different spacing in the prompts for languages such Taiwanese mandarin or Chinese mandarin [524].

C. Computational Requirements

Pre-training LLMs requires significant computational costs which can be very expensive, both financially and environmentally. Millions of dollars are spent in training these LLMs with thousands of compute hours and energy consumption. These are classified as Red AI [525], referring to models achieving state of the art results due to vast computation. Scaling these models can also be a challenging task due to the number of resources invested to train these LLMs. The concept of Computer Optimal Training [522] was introduced to address this problem for maximizing the training efficiency with respect to the corpus and model size. Model parallelism can also be used to distribute the model and the train it faster.

D. Fine-Tuning LLMs

Fine-Tuning LLMs is a useful technique to train LLMs to custom tasks by training further on these task-specific datasets [48]. However, it requires a high amount of memory and large compute resources to store model gradients, parameters and activations, along with storing these fine-tuned models, limiting its access to a few institutions. Parameter-Efficient Fine-Tuning is a technique that can be used to address this problem which consists of updating a subset of model parameters such as prefix fine-tuning [526], prompt-tuning [527] and adapters [528]. Although techniques like Low-Rank Adaptation (LoRA) [529], or LongLora [530], or QLora [531] can be used to optimize the computation cost, but still computational demands remain a significant barrier for Fine-Tuning LLMs.

E. Inference Latency

High inference latency is one of the major challenges of LLMs which is mainly due to large memory footprints and lack of model parallelism. Several techniques can be used to mitigate this problem. Efficient Attention [532] can be used for accelerating attention through sub quadratic approximations such as multi-attention query or flash attention. Quantization [59] can be used to reduce the large memory footprint by reducing the computational precision of activations and weights. Pruning [533] and cascading [534] are some more techniques that can reduce the inference latency drastically for efficient and seamless responses.

F. Limited Context Length

Limited Context Length is one of the crucial aspects of LLMs, as it is extremely useful for interpretation of different prompts and semantic analysis. Without this contextual information, it can drastically degrade the performance of LLMs. There are several strategies that can be used to address this; Positional Embedding Schemas [535], Efficient Attention [536] and Transformer alternatives. Different Positional Embedding Schemes can help LLMs to generalize well to different prompts which may not exist in the training data. Transient Global [537] and Luna [538] are some efficient attention mechanisms that can process larger context lengths effectively. Recurrent Neural Networks (RNNs) [539] and State Space Models (SSMs) [540] are good alternative for transformer-based approaches and are effective for addressing limiting context length.

G. Knowledge Updating and Refinement

Although LLMs are trained on a large corpus of data, the factual information learned may become outdated over time. Retraining the models is a costly process and is not sustainable. To address this, approaches such as model editing [541] is a technique which uses non-parametric knowledge resources to alter a model's behavior, and preserving model parameters by feeding new weights to modify the model's behavior can be used. However, these approaches are found to have limited generalizability and may only be applicable to a limited model architecture. On this end, web plugins and access to the web can alleviate the knowledge updating problem.

H. Risks of Foundation models

A foundation model refers to a base or core model that serves as the fundamental architecture for various machine learning tasks. In [138], a careful assessment of the risks and benefits of foundation models is done. A review [489] also highlights the potential threats and benefits of foundation models in health and education.

- **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. [542], large pre-trained models designed to mimic natural languages can inadvertently perpetuate unfairness and prejudices. The manifestations of these biases are as follows: (i) Training data bias: Language models typically rely on extensive datasets of human language for training [543]. 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 [544].
  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 [545].
- **Information Hallucination:** Hallucination in Natural Language Generation (NLG) is the generation of text that is nonsensical or unfaithful to the provided source content [546]. 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 [390].
  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 [547]. In particular, LLMs might be biased towards producing more "interesting" or fluent outputs, leading to a higher risk of hallucination [548].

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" [549]. Another is to provide the model with a larger and more diverse dataset, which might reduce the risk of the model making incorrect assumptions [77]. In addition to this, 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 [550]. This, however, requires careful consideration of the data and metrics used.

- **Lack of 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 [551], [552]. 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 [553], [554].
  This complexity makes it exceedingly difficult for humans to understand and interpret the decision-making mechanisms employed by the model [85]. The lack of transparency [555] hinders the ability to gain insights into how specific inputs lead to particular outputs [556]. This, in addition to the intricate architecture of LLMs, often consisting of deep neural networks, exacerbates the challenge of explainability [557]. The numerous layers and complex interactions make it challenging to trace the reasoning process of the model. While techniques such as attention mechanisms [558] 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 [559]. 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 [555]. Without explainability, stakeholders may be reluctant to fully trust and adopt LLMs for sensitive applications.
- **Reasoning Errors:** LLM can make mistakes in logical reasoning [560], either because of ambiguities in the prompt or inherent limitations in its understanding of complex logical operations. LLMs cannot plan, reason, and have limited knowledge and commonsense [561] about the physical world [562]. From a cognitive science perspective, Auto-regressive LLMs at their best can approximate the Wernicke and Broca areas in the brain [563].
- **Prompt Injection, 'Jail Break' Attacks [564], 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. Such vulnerabilities have been discussed in [565], [189] through experiments.
- **Adversarial Attacks:** Adversarial attacks on large language models (LLMs) are a type of security threat that can be used to manipulate or control the output of an LLM. These attacks work by deliberately introducing small changes to the input text, which the LLM then misinterprets and produces incorrect or harmful output [566]. One common type of adversarial attack is called a text injection attack. In this type of attack, the attacker introduces carefully crafted text into the input, which the LLM then interprets as a command. For example, the attacker could inject the text "delete all files" into an LLM that is used to control a computer system. The LLM would then delete all of the files on the system [567]. Visual-prompt based models are also being attacked by these corrupted prompts [568].
- **Behavioral Changes over Time** Chen et. al. [569] investigated the performance of GPT 3.5 and GPT 4 over time, between March 2023 to June 2023, and found that the performance can greatly vary over time. For example, In March, GPT-4 had an accuracy of 84%, but in June, its accuracy dropped to 51%, a decrease of 33%. However, many experts suggest that the performance decrease is due to model drift [570] or prompt drift [571], we need to prompt better for maintaining the performance.
- **Spelling and Counting Errors:** Some specific tasks, like identifying and correcting spelling errors, can be challenging for GPT-4 due to its statistical nature. Another such example are counting errors. Counting error occurs when the model miscounts or misinterprets numerical quantities. For instance, it may provide incorrect calculations or misplace decimal points when performing arithmetic operations, and counting the number of words or characters in long paragraphs [521], [572].

X. Open Questions

In this section, we evaluate the open questions that are faced by AI researchers from technical, usage and philosophical standpoints.

A. Environmental and Energy Resources

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 [573]. 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. Moreover, it is expected that the development of newer and advanced version models would need even more significant amounts of water due to their larger data parameters [238]. This concern has been discussed in [574], which presents a method to estimate the water footprint of AI language models and suggests more information transparency in this regard.

Apart from water usage, the training of LLMs demands a considerable amount 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 [575]. 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 [576], [577]. This electricity usage also contributes to indirect water consumption through power generation for data centers located off-site which should be taken into account, leading to carbon emissions [578].

The energy consumption of AI training has significant implications for the environment, particularly in terms of greenhouse gas emissions and climate change [579]. 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 [580]. 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 [581]. One such solution is for data centers to adopt more eco-friendly cooling systems, such as using recycled water or implementing advanced cooling technologies [582].

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 [578]. Another study by Chien et. al [583] found that with models like ChatGPT, inference services dominated the power consumption and the power emissions for one year were equivalent to 25 times the training power of GPT3. They suggested the use of request direction approaches as a promising manner of reducing power consumption in LLMs. Another approach which can be explored is to develop more energy-efficient algorithms and models, which can reduce the amount of energy required to train AI systems.

While, given that generative AI in general has exhibited high emissions during training/inference as mentioned previously, a study found that AI-generated emissions were several times less than the emissions a human would make for writing and illustration tasks [584]. This finding indicates to the need to approach this problem in a more nuanced manner.

B. Ethical Considerations

Inadvertently, LLMs may perpetuate biases inherent in the training data, resulting in outputs that are biased or discriminatory [585] as discussed previously. The challenge lies in identifying and mitigating such biases to ensure fair and equitable treatment across diverse user groups and disciplines [586]. Incorporating robust data authenticity and 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.

1) Humans VS LLMs: Human interactions offer a deep level of empathy, emotional intelligence, and the ability to understand complex nuances in everyday life-situations. Humans responses are not only based on the current situation (prompt), but also considers other factors [195].

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 [587]. 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 and are becoming increasingly autonomous [588]. They are now able to make their own decisions and to take action without human input. In [589], the authors investigatethe differences between AI-generated scientific text and human-written scientific text. They found that AI-generated texts are informative, specific, objective and coherent, but also repetitive, generic, and boring [590].

There is also a need to develop new performance metrics for measuring the intelligence of AI systems, as traditional methods of assessing intelligence, such as IQ tests [317], are not well-suited for AI systems, as they are designed to measure human intelligence [591].

2) Interpretability: Despite their impressive capabilities, LLMs often lack transparency, making it difficult to understand their decision-making process as has been mentioned previously. Enhancing the interpretability of LLMs holds importance for several reasons [592]. 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.

3) Data Efficiency: Data efficiency refers to the efficient use of training data for developing LLMs [204]. LLMs are typically trained using extremely large amounts of data to gain a performance that is acceptable or "human-like". Several techniques which are being explored are data augmentation [593] and data selection [594], knowledge distillation [595], transfer learning [596], meta-learning [597], and others [598].

4) Training data contamination from AI-generated content: Data sources for training large models are typically scraped from the internet. With the increasing popularity of generative AI, it is possible that data present on the internet will have a significant 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 [599] that can determine content generated by AI before passing it through the model during the training process. There is a need to develop a dependable mechanism [497] to perform this task and retain the integrity of data.

5) The future as we perceive it: 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 [600], [601].

- **Source of Inspiration, Suggestions:** LLMs can serve as an invaluable source of inspiration and suggestions, helping users brainstorm ideas [602], create content [603], and make decisions [604].
- **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 [605], [606].
- **Artificial General Intelligence - AGI** Artificial general intelligence (AGI [607]) is a hypothetical type of artificial intelligence that would have the ability to learn and perform any intellectual task. In [312], GPT-4 is found to have sparks of artificial general intelligence. GPT-4 is able to perform a variety of tasks; such as solving math problems, writing creative contents, writing poems and poetry [608] and answering questions in an informative way.
  However, in our opinion, realizing the dream of AGI is still far away, despite of the rapid progress in the LLMs development. The key challenges include; understanding natural intelligence [609], developing adaptable fully autonomous models [610], and being safe and reliable with the understanding of the physical world [611], [612].
- **Democratizing AI** Democratizing AI [613] 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.
  No-code AI platforms [614] may also assist in democratizing AI initiatives [615], by providing a user-friendly interface that allows users to build and deploy ML models without any coding experience [616]. *No-code AI* can be used to leverage machine learning operations (MLOps) [617], to ensure models are deployed and managed effectively in production. 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; such as generating images, chatting with pdf files, and also discussed GPT plug-ins. A comparision of popular chatbots; such as, ChatGPT, Bard, and Bing Chat is also provided.

We particularly explored the applications of LLMs in medicine, engineering, education, finance, media, law, and the entertainment industry. A list of popular LLM-based open-source applications for a variety of tasks is also presented. 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, 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. A thorough comparison between popular chatbots has been provided as well.

Furthermore, the survey has also highlighted the importance of the safe and ethical use of AI tools like ChatGPT and others. 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 other industries is critical, as it enables these tools to effectively support and enhance human endeavors while upholding the values of integrity, privacy, and fairness. In our opinion, A technology X can replace a technology Y on a task Z, and can also help increase the productivity of humans on several tasks. LLMs have a great potential to transform many fields and bring positive impact on humans and society .

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 manner.

Declaration of Interest

The authors have no conflicts of interest to declare.

References

[1] K. S. Jones, "Natural language processing: a historical review," *Current issues in computational linguistics: in honour of Don Walker*, pp. 3–16, 1994.

[2] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[3] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 6, pp. 2515–2528, 2022.

[4] D. Nozza, F. Bianchi, D. Hovy, *et al.*, "Honest: Measuring hurtful sentence completion in language models," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021.

[5] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, 2021.

[6] M. Soam and S. Thakur, "Next word prediction using deep learning: A comparative study," in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 653–658, IEEE, 2022.

[7] S. Diao, R. Xu, H. Su, Y. Jiang, Y. Song, and T. Zhang, "Taming pre-trained language models with n-gram representations for low-resource domain adaptation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3336–3349, 2021.

[8] P. F. Brown, V. J. Della Pietra, P. V. Desouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–480, 1992.

[9] N. Omar and Q. Al-Tashi, "Arabic nested noun compound extraction based on linguistic features and statistical measures," *GEMA Online Journal of Language Studies*, vol. 18, no. 2, pp. 93–107, 2018.

[10] B. Rawat, A. S. Bist, U. Rahardja, Q. Aini, and Y. P. A. Sanjaya, "Recent deep learning based nlp techniques for chatbot development: An exhaustive survey," in *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–4, IEEE, 2022.

[11] Q. Lhoest, A. V. del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, *et al.*, "Datasets: A community library for natural language processing," *arXiv preprint arXiv:2109.02846*, 2021.

[12] O. Sharir, B. Peleg, and Y. Shoham, "The cost of training nlp models: A concise overview," *arXiv preprint arXiv:2004.08900*, 2020.

[13] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[14] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *arXiv preprint arXiv:2206.04615*, 2022.

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] D. Luitse and W. Denkena, "The great transformer: Examining the role of large language models in the political economy of ai," *Big Data & Society*, vol. 8, no. 2, p. 20539517211047734, 2021.

[18] Z. Dong, T. Tang, L. Li, and W. X. Zhao, "A survey on long text modeling with transformers," *arXiv preprint arXiv:2302.14502*, 2023.

[19] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, vol. 6, no. 1, pp. 1–38, 2019.

[20] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundational models defining a new era in vision: A survey and outlook," *arXiv preprint arXiv:2307.13721*, 2023.

[21] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[22] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris, *et al.*, "Avlnet: Learning audio-visual language representations from instructional videos," *arXiv preprint arXiv:2006.09199*, 2020.

[23] Y. Zhao, Z. Lin, D. Zhou, Z. Huang, J. Feng, and B. Kang, "Bubogpt: Enabling visual grounding in multi-modal llms," *arXiv preprint arXiv:2307.08581*, 2023.

[24] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.

[25] N. Pappas and T. Meyer, "A survey on language modeling using neural networks," tech. rep., Idiap, 2012.

[26] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech communication*, vol. 42, no. 1, pp. 93–108, 2004.

[27] J. Lafferty and C. Zhai, "Probabilistic relevance models based on document and query generation," *Language modeling for information retrieval*, pp. 1–10, 2003.

[28] V. A. Petrushin, "Hidden markov models: Fundamentals and applications," in *Online Symposium for Electronics Engineer*, 2000.

[29] S. Khudanpur and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling," *Computer Speech & Language*, vol. 14, no. 4, pp. 355–372, 2000.

[30] H. Wang, J. He, X. Zhang, and S. Liu, "A short text classification method based on n-gram and cnn," *Chinese Journal of Electronics*, vol. 29, no. 2, pp. 248–254, 2020.

[31] 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.

[32] E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep neural network language models," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pp. 20–28, 2012.

[33] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[34] F. Alva-Manchego, C. Scarton, and L. Specia, "Data-driven sentence simplification: Survey and benchmark," *Computational Linguistics*, vol. 46, no. 1, pp. 135–187, 2020.

[35] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021.

[36] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[37] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, no. 1, pp. 1–24, 2015.

[38] M. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pp. 1–5, IEEE, 2013.

[39] A. Go, L. Huang, R. Bhayani, *et al.*, "Twitter sentiment analysis," *Entropy*, vol. 17, p. 252, 2009.

[40] L. Deng and Y. Liu, "A joint introduction to natural language processing and to deep learning," *Deep learning in natural language processing*, pp. 1–22, 2018.

[41] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[42] 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.

[43] S. Hochreiter, "Recurrent neural net learning and vanishing gradient," *International Journal Of Uncertainity, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 107–116, 1998.

[44] S. Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies," *Advances in neural information processing systems*, vol. 8, 1995.

[45] 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.

[46] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[47] B. Ghojogh and A. Ghodsi, "Attention mechanism, transformers, bert, and gpt: tutorial and survey," 2020.

[48] 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.

[49] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," *arXiv preprint arXiv:2003.07278*, 2020.

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

[51] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: a brief survey," *arXiv preprint arXiv:1707.02268*, 2017.

[52] Y. Ge, W. Hua, J. Ji, J. Tan, S. Xu, and Y. Zhang, "Openagi: When llm meets domain experts," *arXiv preprint arXiv:2304.04370*, 2023.

[53] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Llama 2: Early adopters' utilization of meta's new open-source pretrained model," 2023.

[54] A. Baladn, I. Sastre, L. Chiruzzo, and A. Ros, "Retuyt-inco at bea 2023 shared task: Tuning open-source llms for generating teacher responses," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 756–765, 2023.

[55] J. J. Nay, "Large language models as fiduciaries: A case study toward robustly communicating with artificial intelligence through legal standards," 2023.

[56] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[57] T. Y. Zhuo, Z. Li, Y. Huang, Y.-F. Li, W. Wang, G. Haffari, and F. Shiri, "On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex," *arXiv preprint arXiv:2301.12868*, 2023.

[58] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023.

[59] Z. Yao, R. Yazdani Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27168–27183, 2022.

[60] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[61] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 Passes the Bar Exam," March 2023.

[62] A. Byrd, "Truth-telling: Critical inquiries on llms and the corpus texts that train them.," *Composition Studies*, vol. 51, no. 1, pp. 135–142, 2023.

[63] X. Zhang, B. Yu, H. Yu, Y. Lv, T. Liu, F. Huang, H. Xu, and Y. Li, "Wider and deeper llm networks are fairer llm evaluators," *arXiv preprint arXiv:2308.01862*, 2023.

[64] I. Yildirim and L. Paul, "From task structures to world models: What do llms know?," *arXiv preprint arXiv:2310.04276*, 2023.

[65] H. Jin, X. Han, J. Yang, Z. Jiang, C.-Y. Chang, and X. Hu, "Growlength: Accelerating llms pretraining by progressively growing training length," *arXiv preprint arXiv:2310.00576*, 2023.

[66] R. V. P. Marcel, B. E. M. Fernando, and Y. V. J. Roberto, "A brief history of the artificial intelligence: chatgpt: The evolution of gpt," in *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1–5, IEEE, 2023.

[67] E. Y. Chang, "Examining gpt-4: Capabilities, implications, and future directions," 2023.

[68] M. Zhang and J. Li, "A commentary of gpt-3 in mit technology review 2021," *Fundamental Research*, vol. 1, no. 6, pp. 831–833, 2021.

[69] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[70] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[71] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, pp. 1–11, 2023.

[72] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, *et al.*, "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, 2023.

[73] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, and Q. Wang, "Software testing with large language model: Survey, landscape, and vision," *arXiv preprint arXiv:2307.07221*, 2023.

[74] F. F. Xu, U. Alon, G. Neubig, and V. J. Hellendoorn, "A systematic evaluation of large language models of code," in *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pp. 1–10, 2022.

[75] J. Cabrera, M. S. Loyola, I. Magaña, and R. Rojas, "Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots," in *International Work-Conference on Bioinformatics and Biomedical Engineering*, pp. 313–326, Springer, 2023.

[76] A. Creswell, M. Shanahan, and I. Higgins, "Selection-inference: Exploiting large language models for interpretable logical reasoning," *arXiv preprint arXiv:2205.09712*, 2022.

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

[78] K. Tirumala, D. Simig, A. Aghajanyan, and A. S. Morcos, "D4: Improving llm pretraining via document de-duplication and diversification," *arXiv preprint arXiv:2308.12284*, 2023.

[79] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[80] 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.

[81] A. Sudmann, "On the media-political dimension of artificial intelligence: Deep learning as a black box and openai," *Digital Culture & Society*, vol. 4, no. 1, pp. 181–200, 2018.

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

[83] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[84] 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.

[85] 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.

[86] 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*.

[87] 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.

[88] 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.

[89] Q. Al-Tashi, M. B. Saad, A. Sheshadri, C. C. Wu, J. Y. Chang, B. Al-Lazikani, C. Gibbons, N. I. Vokes, J. Zhang, J. J. Lee, *et al.*, "Swarmdeepsurv: swarm intelligence advances deep survival network for prognostic radiomics signatures in four solid cancers," *Patterns*.

[90] S. Mohamadi, G. Mujtaba, N. Le, G. Doretto, and D. A. Adjeroh, "Chatgpt in the age of generative ai and large language models: A concise survey," *arXiv preprint arXiv:2307.04251*, 2023.

[91] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[92] 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.

[93] 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.

[94] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, *et al.*, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *BioRxiv*, vol. 2022, p. 500902, 2022.

[95] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher, "Progen: Language modeling for protein generation," *arXiv preprint arXiv:2004.03497*, 2020.

[96] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[97] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[98] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Pretrained language models for text generation: A survey," *arXiv preprint arXiv:2201.05273*, 2022.

[99] 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.

[100] 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*, 2023.

[101] B. Li, K. Mellou, B. Zhang, J. Pathuri, and I. Menache, "Large language models for supply chain optimization," *arXiv preprint arXiv:2307.03875*, 2023.

[102] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large language models for telecom: The next big thing?," *arXiv preprint arXiv:2306.10249*, 2023.

[103] M. Chen and T. et. al., "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[104] S. Salman, J. A. Shamsi, and R. Qureshi, "Deep fake generation and detection: Issues, challenges, and solutions," *IT Professional*, vol. 25, no. 1, pp. 52–59, 2023.

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

[106] R. Qureshi, M. Irfan, T. M. Gondal, S. Khan, J. Wu, M. U. Hadi, J. Heymach, X. Le, H. Yan, and T. Alam, "Ai in drug discovery and its clinical relevance," *Heliyon*, 2023.

[107] O. B. Shoham and N. Rappoport, "Cpllm: Clinical prediction with large language models," *arXiv preprint arXiv:2309.11295*, 2023.

[108] Q. Al-Tashi, M. B. Saad, A. Muneer, R. Qureshi, S. Mirjalili, A. Sheshadri, X. Le, N. I. Vokes, J. Zhang, and J. Wu, "Machine learning models for the identification of prognostic and predictive cancer biomarkers: A systematic review," *International journal of molecular sciences*, vol. 24, no. 9, p. 7781, 2023.

[109] 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.

[110] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, *et al.*, "A survey on large language model based autonomous agents," *arXiv preprint arXiv:2308.11432*, 2023.

[111] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, and N. Zhang, "Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities," *arXiv preprint arXiv:2305.13168*, 2023.

[112] L. Huynh, J. Hong, A. Mian, H. Suzuki, Y. Wu, and S. Camtepe, "Quantum-inspired machine learning: a survey," *arXiv preprint arXiv:2308.11269*, 2023.

[113] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative ai at work," tech. rep., National Bureau of Economic Research, 2023.

[114] P. Samuelson, "Generative ai meets copyright," *Science*, vol. 381, no. 6654, pp. 158–161, 2023.

[115] I. Chiang, *Unleashing the Power of Generative AI: The Race for Advancement and the Global Ramifications*. PhD thesis, Massachusetts Institute of Technology, 2023.

[116] S. Wang, S. Menon, T. Long, K. Henderson, D. Li, K. Crowston, M. Hansen, J. V. Nickerson, and L. B. Chilton, "Reelframer: Co-creating news reels on social media with generative ai," *arXiv preprint arXiv:2304.09653*, 2023.

[117] S. Mayahi and M. Vidrih, "The impact of generative ai on the future of visual content marketing," *arXiv preprint arXiv:2211.12660*, 2022.

[118] S.-C. Chen, "Multimedia research toward the metaverse," *IEEE MultiMedia*, vol. 29, no. 1, pp. 125–127, 2022.

[119] A. Zentner, "Applied innovation: Artificial intelligence in higher education," *Available at SSRN 4314180*, 2022.

[120] J. Sun, Q. V. Liao, M. Muller, M. Agarwal, S. Houde, K. Talamadupula, and J. D. Weisz, "Investigating explainability of generative ai for code through scenario-based design," in *27th International Conference on Intelligent User Interfaces*, pp. 212–228, 2022.

[121] J. Morley, N. J. DeVito, and J. Zhang, "Generative ai for medical research," 2023.

[122] P. Ghimire, K. Kim, and M. Acharya, "Generative ai in the construction industry: Opportunities & challenges," *arXiv preprint arXiv:2310.04427*, 2023.

[123] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, and B. Wang, "scgpt: Towards building a foundation model for single-cell multi-omics using generative ai," *bioRxiv*, pp. 2023–04, 2023.

[124] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, pp. 159–190, 2006.

[125] A. Pérez-Suárez, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "A review of conceptual clustering algorithms," *Artificial Intelligence Review*, vol. 52, pp. 1267–1296, 2019.

[126] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[127] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: a review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[128] A. A. de Hond, A. M. Leeuwenberg, L. Hooft, I. M. Kant, S. W. Nijman, H. J. van Os, J. J. Aardoom, T. P. Debray, E. Schuit, M. van Smeden, *et al.*, "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review," *NPJ digital medicine*, vol. 5, no. 1, p. 2, 2022.

[129] C. Zhang, C. Zhang, S. Zheng, Y. Qiao, C. Li, M. Zhang, S. K. Dam, C. M. Thwal, Y. L. Tun, L. L. Huy, *et al.*, "A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need?," *arXiv preprint arXiv:2303.11717*, 2023.

[130] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai," *arXiv preprint arXiv:2303.13336*, vol. 2, 2023.

[131] L. Wang, W. Chen, W. Yang, F. Bi, and F. R. Yu, "A state-of-the-art review on image synthesis with generative adversarial networks," *IEEE Access*, vol. 8, pp. 63514–63537, 2020.

[132] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: a review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–25, 2022.

[133] 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.

[134] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[135] S. Hong, J. Seo, S. Hong, H. Shin, and S. Kim, "Large language models are frame-level directors for zero-shot text-to-video generation," *arXiv preprint arXiv:2305.14330*, 2023.

[136] Ö. AYDIN and E. KARAARSLAN, "Is chatgpt leading generative ai? what is beyond expectations," *What is Beyond Expectations*, 2023.

[137] B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo, *et al.*, "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers," *arXiv preprint arXiv:2109.04650*, 2021.

[138] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[139] Y. Yuan, "On the power of foundation models," in *International Conference on Machine Learning*, pp. 40519–40530, PMLR, 2023.

[140] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[141] A. Jo, "The promise and peril of generative ai," *Nature*, vol. 614, no. 1, pp. 214–216, 2023.

[142] H. Bansal, K. Gopalakrishnan, S. Dingliwal, S. Bodapati, K. Kirchhoff, and D. Roth, "Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale," *arXiv preprint arXiv:2212.09095*, 2022.

[143] M. Mariani, "Generative artificial intelligence and innovation: Conceptual foundations," *Available at SSRN 4249382*, 2022.

[144] 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.

[145] 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.

[146] 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.

[147] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[148] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[149] 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.

[150] 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.

[151] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: introduction and outlook," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 588–598, 2017.

[152] J. N. Kather, N. Ghaffari Laleh, S. Foersch, and D. Truhn, "Medical domain knowledge in domain-agnostic generative ai," *NPJ digital medicine*, vol. 5, no. 1, p. 90, 2022.

[153] I. Solaiman, "The gradient of generative ai release: Methods and considerations," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 111–122, 2023.

[154] S. Tan, Y. Shen, and B. Zhou, "Improving the fairness of deep generative models without retraining," *arXiv preprint arXiv:2012.04842*, 2020.

[155] K. Wach, C. D. Duong, J. Ejdys, R. Kazlauskaitė, P. Korzynski, G. Mazurek, J. Paliszkiewicz, and E. Ziemba, "The dark side of generative artificial intelligence: A critical analysis of controversies and risks of chatgpt," *Entrepreneurial Business and Economics Review*, vol. 11, no. 2, pp. 7–24, 2023.

[156] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[157] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 787–792, IEEE, 2017.

[158] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[159] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "Clustergan: Latent space clustering in generative adversarial networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 4610–4617, 2019.

[160] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, "Chunked autoregressive gan for conditional waveform synthesis," *arXiv preprint arXiv:2110.10139*, 2021.

[161] S. Kaushik, A. Choudhury, S. Natarajan, L. A. Pickett, and V. Dutt, "Medicine expenditure prediction via a variance-based generative adversarial network," *IEEE Access*, vol. 8, pp. 110947–110958, 2020.

[162] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4773–4784, 2020.

[163] N. Geneva and N. Zabaras, "Multi-fidelity generative deep learning turbulent flows," *arXiv preprint arXiv:2006.04731*, 2020.

[164] J. D. Cohen, S. M. McClure, and A. J. Yu, "Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1481, pp. 933–942, 2007.

[165] D. Xu, F. Zhu, Q. Liu, and P. Zhao, "Improving exploration efficiency of deep reinforcement learning through samples produced by generative model," *Expert Systems with Applications*, vol. 185, p. 115680, 2021.

[166] P. Dhoni, "Exploring the synergy between generative ai, data and analytics in the modern age," 2023.

[167] G. Vigliensoni, P. Perry, R. Fiebrink, *et al.*, "A small-data mindset for generative ai creative work," 2022.

[168] Y. Kossale, M. Airaj, and A. Darouichi, "Mode collapse in generative adversarial networks: An overview," in *2022 8th International Conference on Optimization and Applications (ICOA)*, pp. 1–6, IEEE, 2022.

[169] Y. Ding, N. Mishra, and H. Hoffmann, "Generative and multi-phase learning for computer systems optimization," in *Proceedings of the 46th International Symposium on Computer Architecture*, pp. 39–52, 2019.

[170] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges," *Future Internet*, vol. 15, no. 8, p. 260, 2023.

[171] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. Smagt, "Metrics for deep generative models," in *International Conference on Artificial Intelligence and Statistics*, pp. 1540–1550, PMLR, 2018.

[172] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[173] A. Obukhov and M. Krasnyanskiy, "Quality assessment method for gan based on modified metrics inception score and fréchet inception distance," in *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4*, pp. 102–114, Springer, 2020.

[174] A. Verine, B. Negrevergne, M. S. Pydi, and Y. Chevaleyre, "Precision-recall divergence optimization for generative modeling with gans and normalizing flows," *arXiv preprint arXiv:2305.18910*, 2023.

[175] R. Kansal, J. Duarte, H. Su, B. Orzari, T. Tomei, M. Pierini, M. Touranakou, D. Gunopulos, *et al.*, "Particle cloud generation with message passing generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23858–23871, 2021.

[176] A. Borji, "Pros and cons of gan evaluation measures: New developments," *Computer Vision and Image Understanding*, vol. 215, p. 103329, 2022.

[177] DeepLearning.AI, "Generative AI with LLMs." https://www.deeplearning.ai/courses/generative-ai-with-llms/, n.d. Coursera.

[178] A. Schmidt, "Speeding up the engineering of interactive systems with generative ai," in *Companion Proceedings of the 2023 ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pp. 7–8, 2023.

[179] H. Muse, S. Bulathwela, and E. Yilmaz, "Pre-training with scientific text improves educational question generation (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 16288–16289, 2023.

[180] D. Foster, *Generative deep learning*. " O'Reilly Media, Inc.", 2022.

[181] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu, "Aligning large language models with human: A survey," *arXiv preprint arXiv:2307.12966*, 2023.

[182] R. Zhong, K. Lee, Z. Zhang, and D. Klein, "Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections," *arXiv preprint arXiv:2104.04670*, 2021.

[183] H. Dang, L. Mecke, F. Lehmann, S. Goller, and D. Buschek, "How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models," *arXiv preprint arXiv:2209.01390*, 2022.

[184] F. Song, B. Yu, M. Li, H. Yu, F. Huang, Y. Li, and H. Wang, "Preference ranking optimization for human alignment," *arXiv preprint arXiv:2306.17492*, 2023.

[185] H. Liu, Z. Teng, L. Cui, C. Zhang, Q. Zhou, and Y. Zhang, "Logicot: Logical chain-of-thought instruction-tuning data collection with gpt-4," *arXiv preprint arXiv:2305.12147*, 2023.

[186] J. Oppenlaender, "Prompt engineering for text-based generative art," *arXiv preprint arXiv:2204.13988*, 2022.

[187] N. Ratner, Y. Levine, Y. Belinkov, O. Ram, I. Magar, O. Abend, E. Karpas, A. Shashua, K. Leyton-Brown, and Y. Shoham, "Parallel context windows for large language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6383–6402, 2023.

[188] T. Xue, Z. Wang, and H. Ji, "Parameter-efficient tuning helps language model alignment," *arXiv preprint arXiv:2310.00819*, 2023.

[189] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and Y. Liu, "Jailbreaking chatgpt via prompt engineering: An empirical study," *arXiv preprint arXiv:2305.13860*, 2023.

[190] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou, "Towards interpretable mental health analysis with chatgpt," 2023.

[191] J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, "Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design," *arXiv preprint arXiv:2303.07839*, 2023.

[192] C. Liu, X. Bao, H. Zhang, N. Zhang, H. Hu, X. Zhang, and M. Yan, "Improving chatgpt prompt for code generation," *arXiv preprint arXiv:2305.08360*, 2023.

[193] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530, IEEE, 2023.

[194] S. K. K. Santu and D. Feng, "Teler: A general taxonomy of llm prompts for benchmarking complex tasks," *arXiv preprint arXiv:2305.11430*, 2023.

[195] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, "Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2023.

[196] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[197] S. Diao, P. Wang, Y. Lin, and T. Zhang, "Active prompting with chain-of-thought for large language models," *arXiv preprint arXiv:2302.12246*, 2023.

[198] P. Denny, J. Leinonen, J. Prather, A. Luxton-Reilly, T. Amarouche, B. A. Becker, and B. N. Reeves, "Promptly: Using prompt problems to teach learners how to effectively utilize ai code generators," *arXiv preprint arXiv:2307.16364*, 2023.

[199] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex, "Planning with large language models via corrective re-prompting," *arXiv preprint arXiv:2211.09935*, 2022.

[200] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey for in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[201] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.

[202] F. Liu, J. M. Eisenschlos, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, W. Chen, N. Collier, and Y. Altun, "Deplot: One-shot visual language reasoning by plot-to-table translation," *arXiv preprint arXiv:2212.10505*, 2022.

[203] X. Liu, D. McDuff, G. Kovacs, I. Galatzer-Levy, J. Sunshine, J. Zhan, M.-Z. Poh, S. Liao, P. Di Achille, and S. Patel, "Large language models are few-shot health learners," *arXiv preprint arXiv:2305.15525*, 2023.

[204] Z. Hu, Y. Lan, L. Wang, W. Xu, E.-P. Lim, R. K.-W. Lee, L. Bing, and S. Poria, "Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models," *arXiv preprint arXiv:2304.01933*, 2023.

[205] D. Miyake, A. Iohara, Y. Saito, and T. Tanaka, "Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models," *arXiv preprint arXiv:2305.16807*, 2023.

[206] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *European Conference on Computer Vision*, pp. 423–439, Springer, 2022.

[207] AUTOMATIC1111, "Negative-prompt." https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt, 2022. Accessed on August 1, 2023.

[208] F. Ma, C. Zhang, L. Ren, J. Wang, Q. Wang, W. Wu, X. Quan, and D. Song, "Xprompt: Exploring the extreme of prompt tuning," *arXiv preprint arXiv:2210.04457*, 2022.

[209] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

[210] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, *et al.*, "Muse: Text-to-image generation via masked generative transformers," *arXiv preprint arXiv:2301.00704*, 2023.

[211] A. Chen, Y. Yao, P.-Y. Chen, Y. Zhang, and S. Liu, "Understanding and improving visual prompting: A label-mapping perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19133–19143, 2023.

[212] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. Efros, "Visual prompting via image inpainting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25005–25017, 2022.

[213] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*, pp. 709–727, Springer, 2022.

[214] T. Chakrabarty, A. Saakyan, O. Winn, A. Panagopoulou, Y. Yang, M. Apidianaki, and S. Muresan, "I spy a metaphor: Large language models and diffusion models co-create visual metaphors," *arXiv preprint arXiv:2305.14724*, 2023.

[215] R. Volum, S. Rao, M. Xu, G. A. DesGarennes, C. Brockett, B. Van Durme, O. Deng, A. Malhotra, and B. Dolan, "Craft an iron sword: Dynamically generating interactive game characters by prompting large language models tuned on code," in *The Third Wordplay: When Language Meets Games Workshop*, 2022.

[216] D. Hegde, J. M. J. Valanarasu, and V. M. Patel, "Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition," *arXiv preprint arXiv:2303.11313*, 2023.

[217] 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.

[218] L. Pan, A. Albalak, X. Wang, and W. Y. Wang, "Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning," *arXiv preprint arXiv:2305.12295*, 2023.

[219] 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.

[220] 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.

[221] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[222] M. Setnes, R. Babuska, and H. B. Verbruggen, "Rule-based modeling: Precision and transparency," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 28, no. 1, pp. 165–169, 1998.

[223] M. Wu, F. Liu, and T. Cohn, "Evaluating the utility of hand-crafted features in sequence labelling," *arXiv preprint arXiv:1808.09075*, 2018.

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

[225] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, pp. 177–196, 2001.

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

[227] 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.

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

[229] 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.

[230] 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.

[231] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[232] 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.

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

[234] 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.

[235] R. Dale, "Gpt-3: What's it good for?," *Natural Language Engineering*, vol. 27, no. 1, pp. 113–118, 2021.

[236] 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.

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

[238]

[239] M. R. King, "Can bard, google's experimental chatbot based on the lamda large language model, help to analyze the gender and racial diversity of authors in your cited scientific references?," *Cellular and Molecular Bioengineering*, vol. 16, no. 2, pp. 175–179, 2023.

[240] C. Khatri, B. Hedayatnia, A. Venkatesh, J. Nunn, Y. Pan, Q. Liu, H. Song, A. Gottardi, S. Kwatra, S. Pancholi, *et al.*, "Advancing the state of the art in open domain dialog systems through the alexa prize," *arXiv preprint arXiv:1812.10757*, 2018.

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

[242] X. Liu, H. Cheng, P. He, W. Chen, Y. Wang, H. Poon, and J. Gao, "Adversarial training for large neural language models," *arXiv preprint arXiv:2004.08994*, 2020.

[243] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," *arXiv preprint arXiv:1312.3005*, 2013.

[244] S. Biderman and S. et. al., "Pythia: A suite for analyzing large language models across training and scaling," *arXiv preprint arXiv:2304.01373*, 2023.

[245] 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*.

[246] 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*.

[247] 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.

[248] P. Banerjee and H. Han, "Language modeling approaches to information retrieval," 2009.

[249] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[250] 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.

[251] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade: Second Edition*, pp. 421–436, Springer, 2012.

[252] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[253] 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.

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

[255] Y. Sun, Y. Zheng, C. Hao, and H. Qiu, "Nsp-bert: A prompt-based zero-shot learner through an original pre-training task–next sentence prediction," *arXiv e-prints*, pp. arXiv–2109, 2021.

[256] 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*.

[257] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," in *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pp. 194–206, Springer, 2019.

[258] 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.

[259] 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.

[260] 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.

[261] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Llm. int8 (): 8-bit matrix multiplication for transformers at scale," *arXiv preprint arXiv:2208.07339*, 2022.

[262] Z. Liu, J. Wang, T. Dao, T. Zhou, B. Yuan, Z. Song, A. Shrivastava, C. Zhang, Y. Tian, C. Re, *et al.*, "Deja vu: Contextual sparsity for efficient llms at inference time," in *International Conference on Machine Learning*, pp. 22137–22176, PMLR, 2023.

[263] 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.

[264] H. Ye, Z. Chen, D.-H. Wang, and B. Davison, "Pretrained generalized autoregressive model with adaptive probabilistic label clusters for extreme multi-label text classification," in *International Conference on Machine Learning*, pp. 10809–10819, PMLR, 2020.

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

[266] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[267] M. Hobbhahn, T. Lieberum, and D. Seiler, "Investigating causal understanding in llms," in *NeurIPS ML Safety Workshop*, 2022.

[268] 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.

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

[270] 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.

[271] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *International Conference on Machine Learning*, pp. 1931–1942, PMLR, 2021.

[272] 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.

[273] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.

[274] D. Su, Y. Xu, G. I. Winata, P. Xu, H. Kim, Z. Liu, and P. Fung, "Generalizing question answering system with pre-trained language model fine-tuning," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 203–211, 2019.

[275] N. Pawlowski, J. Vaughan, J. Jennings, and C. Zhang, "Answering causal questions with augmented llms," 2023.

[276] Y. Zhuang, Y. Yu, K. Wang, H. Sun, and C. Zhang, "Toolqa: A dataset for llm question answering with external tools," *arXiv preprint arXiv:2306.13304*, 2023.

[277] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, pp. 1–9, 2023.

[278] A. Piñeiro-Martín, C. García-Mateo, L. Docío-Fernández, and M. d. C. López-Pérez, "Ethical challenges in the development of virtual assistants powered by large language models," *Electronics*, vol. 12, no. 14, p. 3170, 2023.

[279] A. Lazaridou, E. Gribovskaya, W. Stokowiec, and N. Grigorev, "Internet-augmented language models through few-shot prompting for open-domain question answering," *arXiv preprint arXiv:2203.05115*, 2022.

[280] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[281] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," *arXiv preprint arXiv:2006.14799*, 2020.

[282] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," *arXiv preprint arXiv:1912.02164*, 2019.

[283] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, "A generative model for category text generation," *Information Sciences*, vol. 450, pp. 301–315, 2018.

[284] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, and Z. Tu, "Document-level machine translation with large language models," *arXiv preprint arXiv:2304.02210*, 2023.

[285] H. Huang, S. Wu, X. Liang, B. Wang, Y. Shi, P. Wu, M. Yang, and T. Zhao, "Towards making the most of llm for translation quality estimation," in *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 375–386, Springer, 2023.

[286] C. Lyu, J. Xu, and L. Wang, "New trends in machine translation using large language models: Case examples with chatgpt," *arXiv preprint arXiv:2305.01181*, 2023.

[287] B. Zhang, B. Haddow, and A. Birch, "Prompting large language model for machine translation: A case study," *arXiv preprint arXiv:2301.07069*, 2023.

[288] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, "Text classification via large language models," *arXiv preprint arXiv:2305.08377*, 2023.

[289] R. Song, Z. Liu, X. Chen, H. An, Z. Zhang, X. Wang, and H. Xu, "Label prompt for multi-label text classification," *Applied Intelligence*, vol. 53, no. 8, pp. 8761–8775, 2023.

[290] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," *arXiv preprint arXiv:2305.15005*, 2023.

[291] M. Labonne and S. Moran, "Spam-t5: Benchmarking large language models for few-shot email spam detection," *arXiv preprint arXiv:2304.01238*, 2023.

[292] S. S. Mullick, M. Bhambhani, S. Sinha, A. Mathur, S. Gupta, and J. Shah, "Content moderation for evolving policies using binary question answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 561–573, 2023.

[293] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, *et al.*, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *arXiv preprint arXiv:2302.12813*, 2023.

[294] R. C. Fernandez, A. J. Elmore, M. J. Franklin, S. Krishnan, and C. Tan, "How large language models will disrupt data management,"

Proceedings of the VLDB Endowment, vol. 16, no. 11, pp. 3302–3309, 2023.

[295] S. Yu, C. Fang, Y. Ling, C. Wu, and Z. Chen, "Llm for test script generation and migration: Challenges, capabilities, and opportunities," *arXiv preprint arXiv:2309.13574*, 2023.

[296] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *arXiv preprint arXiv:2301.13848*, 2023.

[297] C. Shen, L. Cheng, Y. You, and L. Bing, "Are large language models good evaluators for abstractive summarization?," *arXiv preprint arXiv:2305.13091*, 2023.

[298] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *arXiv preprint arXiv:2305.05658*, 2023.

[299] R. Luo, Z. Zhao, M. Yang, J. Dong, M. Qiu, P. Lu, T. Wang, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," *arXiv preprint arXiv:2306.07207*, 2023.

[300] S. Wadhwa, S. Amir, and B. C. Wallace, "Revisiting relation extraction in the era of large language models," *arXiv preprint arXiv:2305.05003*, 2023.

[301] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, *et al.*, "Zero-shot information extraction via chatting with chatgpt," *arXiv preprint arXiv:2302.10205*, 2023.

[302] C. Li, X. Zhang, D. Chrysostomou, and H. Yang, "Tod4ir: A humanised task-oriented dialogue system for industrial robots," *IEEE Access*, vol. 10, pp. 91631–91649, 2022.

[303] J. Deng, H. Sun, Z. Zhang, J. Cheng, and M. Huang, "Recent advances towards safe, responsible, and moral dialogue systems: A survey," *arXiv preprint arXiv:2302.09270*, 2023.

[304] J. Wei, J. Wei, Y. Tay, D. Tran, A. Webson, Y. Lu, X. Chen, H. Liu, D. Huang, D. Zhou, *et al.*, "Larger language models do in-context learning differently," *arXiv preprint arXiv:2303.03846*, 2023.

[305] V. Bhat and P. Bhattacharyya, "Survey: Automatic speech recognition for indian languages,"

[306] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, pp. 28492–28518, PMLR, 2023.

[307] 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.

[308] 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.

[309] 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.

[310] 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.

[311] 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.

[312] 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.

[313] 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.

[314] 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.

[315] 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.

[316] F. C. Kitamura, "Chatgpt is shaping the future of medical writing but still requires human judgment," 2023.

[317] 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.

[318] 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.

[319] 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.

[320] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios," *Journal of Medical Systems*, vol. 47, no. 1, p. 33, 2023.

[321] S. K. Karn, R. Ghosh, O. Farri, *et al.*, "shs-nlp at radsum23: Domain-adaptive pre-training of instruction-tuned llms for radiology report impression generation," *arXiv preprint arXiv:2306.03264*, 2023.

[322] 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.

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

[324] 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.

[325] M. F. Romano, L. C. Shih, I. C. Paschalidis, R. Au, and V. B. Kolachalama, "Large language models in neurology research and future practice," *Neurology*, 2023.

[326] 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.

[327] S. M. Jungmann, T. Klan, S. Kuhn, and F. Jungmann, "Accuracy of a chatbot (ada) in the diagnosis of mental disorders: comparative case study with lay and expert users," *JMIR formative research*, vol. 3, no. 4, p. e13863, 2019.

[328] D. Magalhaes Azevedo and S. Kieffer, "User reception of ai-enabled mhealth apps: The case of babylon health.,"

[329] P. Malik, M. Pathania, V. K. Rathaur, *et al.*, "Overview of artificial intelligence in medicine," *Journal of family medicine and primary care*, vol. 8, no. 7, p. 2328, 2019.

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

[331] J. Ma and B. Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[332] Y. Li, C. Gao, X. Song, X. Wang, Y. Xu, and S. Han, "Druggpt: A gpt-based strategy for designing potential ligands targeting specific proteins," *bioRxiv*, pp. 2023–06, 2023.

[333] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[334] C. Cohn, N. Hutchins, and G. Biswas, "Towards a formative feedback generation agent: Leveraging a human-in-the-loop, chain-of-thought prompting approach with llms to evaluate formative assessment responses in k-12 science.," 2023.

[335] Y. Wu, S. Y. Min, Y. Bisk, R. Salakhutdinov, A. Azaria, Y. Li, T. Mitchell, and S. Prabhumoye, "Plan, eliminate, and track–language models are good teachers for embodied agents," *arXiv preprint arXiv:2305.02412*, 2023.

[336] S. Doveh, A. Arbelle, S. Harary, E. Schwartz, R. Herzig, R. Giryes, R. Feris, R. Panda, S. Ullman, and L. Karlinsky, "Teaching structured vision & language concepts to vision & language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2657–2668, 2023.

[337] S. Saha, P. Hase, and M. Bansal, "Can language models teach weaker agents? teacher explanations improve students via theory of mind," *arXiv preprint arXiv:2306.09299*, 2023.

[338] S. Peng, K. Yuan, L. Gao, and Z. Tang, "Mathbert: A pre-trained model for mathematical formula understanding," *arXiv preprint arXiv:2105.00377*, 2021.

[339] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević, "Practical and ethical challenges of large language models in education: A systematic literature review," *arXiv preprint arXiv:2303.13379*, 2023.

[340] 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.

[341] "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.

[342] 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.

[343] L. Zhu, W. Mou, T. Yang, and R. Chen, "Chatgpt can pass the aha exams: Open-ended questions outperform multiple-choice format," *Resuscitation*, vol. 188, p. 109783, 2023.

[344] 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.

[345] B.-C. Kuo, F. T. Chang, and Z.-E. Bai, "Leveraging llms for adaptive testing and learning in taiwan adaptive learning platform (talp)," 2023.

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

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

[348] T. Soubhari, S. S. Nanda, T. A. Lone, and P. S. Beegam, "Digital hacks, creativity shacks, and academic menace: The ai effect," in *Sustainable Development Goal Advancement Through Digital Innovation in the Service Sector*, pp. 208–232, IGI Global, 2023.

[349] J. Prather, P. Denny, J. Leinonen, B. A. Becker, I. Albluwi, M. Craig, H. Keuning, N. Kiesler, T. Kohn, A. Luxton-Reilly, *et al.*, "The robots are here: Navigating the generative ai revolution in computing education," *arXiv preprint arXiv:2310.00658*, 2023.

[350] H. H. Thorp, "Chatgpt is fun, but not an author," 2023.

[351] C. Stokel-Walker, "Chatgpt listed as author on research papers: many scientists disapprove," *Nature*, vol. 613, no. 7945, pp. 620–621, 2023.

[352] 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.

[353] O. Buruk, "Academic writing with gpt-3.5: Reflections on practices, efficacy and transparency," *arXiv preprint arXiv:2304.11079*, 2023.

[354] M. Dowling and B. Lucey, "Chatgpt for (finance) research: The bananarama conjecture," *Finance Research Letters*, vol. 53, p. 103662, 2023.

[355] X.-Y. Liu, G. Wang, and D. Zha, "Fingpt: Democratizing internet-scale data for financial large language models," *arXiv preprint arXiv:2307.10485*, 2023.

[356] A. Zaremba and E. Demir, "Chatgpt: Unlocking the future of nlp in finance," *Available at SSRN 4323643*, 2023.

[357] A. Lopez-Lira and Y. Tang, "Can chatgpt forecast stock price movements? return predictability and large language models," *arXiv preprint arXiv:2304.07619*, 2023.

[358] Y. Yang, M. C. S. Uy, and A. Huang, "Finbert: A pretrained language model for financial communications," *arXiv preprint arXiv:2006.08097*, 2020.

[359] D. Peskoff and B. M. Stewart, "Credible without credit: Domain experts assess generative language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 427–438, 2023.

[360] S. Agarwal, S. Alok, P. Ghosh, and S. Gupta, "Financial inclusion and alternate credit scoring for the millennials: role of big data and machine learning in fintech," *Business School, National University of Singapore Working Paper, SSRN*, vol. 3507827, 2020.

[361] K. B. Hansen, "The virtue of simplicity: On machine learning models in algorithmic trading," *Big Data & Society*, vol. 7, no. 1, p. 2053951720926558, 2020.

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

[363] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, and H. Wang, "Large language models for software engineering: A systematic literature review," *arXiv preprint arXiv:2308.10620*, 2023.

[364] 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.

[365] 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.

[366] A. Agala, M. Khan, and A. Starr, "Degradation mechanisms associated with metal pipes and the effective impact of ldms and llms in water transport and distribution," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 237, no. 8, pp. 1855–1876, 2023.

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

[368] 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.

[369] J. V. Pavlik, "Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education," *Journalism & Mass Communication Educator*, vol. 78, no. 1, pp. 84–93, 2023.

[370] L. Chan, L. Hogaboam, and R. Cao, "Ai in media and entertainment," in *Applied Artificial Intelligence in Business: Concepts and Cases*, pp. 305–324, Springer, 2022.

[371] R. Lachman and M. Joffe, "Applications of artificial intelligence in media and entertainment," in *Analyzing future applications of AI, sensors, and robotics in society*, pp. 201–220, IGI Global, 2021.

[372] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," *arXiv preprint arXiv:2301.10226*, 2023.

[373] Z. Wang, "Mediagpt: A large language model target chinese media," *arXiv preprint arXiv:2307.10930*, 2023.

[374] J. M. Pérez, D. A. Furman, L. A. Alemany, and F. Luque, "Robertuito: a pre-trained language model for social media text in spanish," *arXiv preprint arXiv:2111.09453*, 2021.

[375] M. Abdulhai, C. Crepy, D. Valter, J. Canny, and N. Jaques, "Moral foundations of large language models," in *AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI*, 2022.

[376] H. Steck, L. Baltrunas, E. Elahi, D. Liang, Y. Raimond, and J. Basilico, "Deep learning for recommender systems: A netflix case study," *AI Magazine*, vol. 42, no. 3, pp. 7–18, 2021.

[377] "Databricks - Media Entertainment Solutions." https://www.databricks.com/solutions/industries/media-and-entertainment. Accessed: Insert date accessed.

[378] J. Kim, K. Xu, and K. Merrill Jr, "Man vs. machine: Human responses to an ai newscaster and the role of social presence," *The Social Science Journal*, pp. 1–13, 2022.

[379] M. Feng, "The development of" ai" synthetic anchor in the context of artificial intelligence," *Highlights in Art and Design*, vol. 2, no. 1, pp. 38–40, 2023.

[380] A. of the article, "Ai-generated news presenter appears in kuwait," *Al Jazeera*, April 2023.

[381] A. of the article, "This is how ai could change the future of journalism," *Sky News*, 2023.

[382] K. Y. Iu and V. M.-Y. Wong, "ChatGPT by OpenAI: The End of Litigation Lawyers?," 2023. Available at SSRN.

[383] M. Ajevski, K. Barker, A. Gilbert, L. Hardie, and F. Ryan, "Chatgpt and the future of legal education and practice," *The Law Teacher*, vol. 0, no. 0, pp. 1–13, 2023.

[384] "The legal ai you've been waiting for." https://casetext.com/cocounsel/. 31 July 2023].

[385] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, "Chatlaw: Open-source legal large language model with integrated external knowledge bases," *arXiv preprint arXiv:2306.16092*, 2023.

[386] J. J. Nay, "Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans," *arXiv preprint*, 2022.

[387] D. Trautmann, A. Petrova, and F. Schilder, "Legal Prompt Engineering for Multilingual Legal Judgement Prediction," *arXiv preprint*, 2022.

[388] J. H. Choi, K. E. Hickman, A. Monahan, and D. B. Schwarcz, "ChatGPT Goes to Law School," *Journal of Legal Education*, January 2023. Forthcoming.

[389] F. Yu, L. Quartey, and F. Schilder, "Legal prompting: Teaching a language model to think like a lawyer," 2022.

[390] J. H. Choi, K. E. Hickman, A. Monahan, and D. B. Schwarcz, "Supra Note 7," 2023. Reference to a previously cited work.

[391] The Guardian, "Two US Lawyers Fined for Submitting Fake Court Citations by ChatGPT," June 2023.

[392] ABC News, "US Lawyer Uses ChatGPT to Research Case with Embarrassing Result," June 2023.

[393] Business Standard, "US Judge Orders Lawyers Not to Use ChatGPT-drafted Content in Court," May 2023.

[394] A. K. Kushwaha and A. K. Kar, "Language model-driven chatbot for business to address marketing and selection of products," in *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2020, Tiruchirappalli, India, December 18–19, 2020, Proceedings, Part I*, pp. 16–28, Springer, 2020.

[395] P. Rivas and L. Zhao, "Marketing with ChatGPT: Navigating the Ethical Terrain of GPT-Based Chatbot Technology," *AI*, vol. 4, pp. 375–384, 2023.

[396] S. Verma, R. Sharma, S. Deb, and D. Maitra, "Artificial intelligence in marketing: Systematic review and future research direction," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100002, 2021.

[397] C. Zielinski, M. Winker, R. Aggarwal, L. Ferris, M. Heinemann, J. Lapeña, S. Pai, and L. Citrome, "Chatbots, ChatGPT, and Scholarly Manuscripts - WAME Recommendations on ChatGPT and Chatbots in Relation to Scholarly Publications," *Afro-Egypt. J. Infect. Endem. Dis.*, vol. 13, pp. 75–79, 2023.

[398] A. F.-B. Sun, Grace H. DNP and R.-B.-C. C. F. Hoelscher, Stephanie H. DNP, "The ChatGPT Storm and What Faculty Can Do," *Nurse Educator*, vol. 48, pp. 119–124, May/June 2023.

[399] L. Ma and B. Sun, "Machine learning and ai in marketing – connecting computing power to human insights," *International Journal of Research in Marketing*, vol. 37, no. 3, pp. 481–504, 2020.

[400] O. Yara, A. Brazheyev, L. Golovko, and V. Bashkatova, "Legal regulation of the use of artificial intelligence: Problems and development prospects," *European Journal of Sustainable Development*, vol. 10, p. 281, Feb. 2021.

[401] M. Stone, E. Aravopoulou, Y. Ekinci, G. Evans, M. Hobbs, A. Labib, P. Laughlin, J. Machtynger, and L. Machtynger, "Artificial Intelligence (AI) in Strategic Marketing Decision-Making: A research agenda," *Bottom Line*, vol. 33, pp. 183–200, 2020.

[402] E. Hermann, "Leveraging Artificial Intelligence in Marketing for Social Good—An Ethical Perspective," *J Bus Ethics*, vol. 179, pp. 43–61, 2022.

[403] K. Jarek and G. Mazurek, "Marketing and artificial intelligence.," *Central European Business Review*, vol. 8, no. 2, 2019.

[404] Z. Liu, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, W. Liu, D. Shen, Q. Li, *et al.*, "Deid-gpt: Zero-shot medical text de-identification by gpt-4," *arXiv preprint arXiv:2303.11032*, 2023.

[405] A. D. Subagja, A. M. A. Ausat, A. R. Sari, M. I. Wanof, and S. Suherlan, "Improving customer service quality in msmes through the use of chatgpt," *Jurnal Minfo Polgan*, vol. 12, no. 2, pp. 380–386, 2023.

[406] K. Howell, G. Christian, P. Fomitchov, G. Kehat, J. Marzulla, L. Rolston, J. Tredup, I. Zimmerman, E. Selfridge, and J. Bradley, "The economic trade-offs of large language models: A case study," *arXiv preprint arXiv:2306.07402*, 2023.

[407] J. Potts, D. W. Allen, C. Berg, and N. Ilyushina, "Large language models reduce agency costs," *Available at SSRN*, 2023.

[408] P. A. Olujimi and A. Ade-Ibijola, "Nlp techniques for automating responses to customer queries: a systematic review," *Discover Artificial Intelligence*, vol. 3, no. 1, p. 20, 2023.

[409] S. Makridakis, F. Petropoulos, and Y. Kang, "Large language models: Their success and impact," *Forecasting*, vol. 5, no. 3, pp. 536–549, 2023.

[410] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang, "Autogen: Enabling next-gen llm applications via multi-agent conversation framework," *arXiv preprint arXiv:2308.08155*, 2023.

[411] N. Morrical, J. Tremblay, Y. Lin, S. Tyree, S. Birchfield, V. Pascucci, and I. Wald, "Nvisii: A scriptable tool for photorealistic image generation," *arXiv preprint arXiv:2105.13962*, 2021.

[412] Y. Li, M. Min, D. Shen, D. Carlson, and L. Carin, "Video generation from text," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[413] G. Franceschelli and M. Musolesi, "On the creativity of large language models," *arXiv preprint arXiv:2304.00008*, 2023.

[414] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in neural information processing systems*, vol. 32, 2019.

[415] 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.

[416] X.-Q. Dao, "Performance comparison of large language models on vnhsge english dataset: Openai chatgpt, microsoft bing chat, and google bard," *arXiv preprint arXiv:2307.02288*, 2023.

[417] 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.

[418] I. Ahmed, M. Kajol, U. Hasan, P. P. Datta, A. Roy, and M. R. Reza, "Chatgpt vs. bard: A comparative study," *UMBC Student Collection*, 2023.

[419] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[420] X. Amatriain, "Transformer models: an introduction and catalog," *arXiv preprint arXiv:2302.07730*, 2023.

[421] Y. Hu, I. Ameer, X. Zuo, X. Peng, Y. Zhou, Z. Li, Y. Li, J. Li, X. Jiang, and H. Xu, "Zero-shot clinical entity recognition using chatgpt," *arXiv preprint arXiv:2303.16416*, 2023.

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

[423] M. Abdullah, A. Madain, and Y. Jararweh, "Chatgpt: Fundamentals, applications and social impacts," in *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 1–8, IEEE, 2022.

[424] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[425] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja, "Lambda architecture for cost-effective batch and speed big data processing," in *2015 IEEE international conference on big data (big data)*, pp. 2785–2792, IEEE, 2015.

[426] D. Xuan-Quy, L. Ngoc-Bich, V. The-Duy, P. Xuan-Dung, N. Bac-Bien, N. Van-Tien, N. Thi-My-Thanh, and N. Hong-Phuoc, "Vnhsge: Vietnamese high school graduation examination dataset for large language models," *arXiv preprint arXiv:2305.12199*, 2023.

[427] H. Trng, "Chatgpt in education-a global and vietnamese research overview," 2023.

[428] S. Bhardwaz and J. Kumar, "An extensive comparative analysis of chatbot technologies - chatgpt, google bard and microsoft bing," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 673–679, 2023.

[429] B. Campello de Souza, A. Serrano de Andrade Neto, and A. Roazzi, "Are the new ais smart enough to steal your job? iq scores for chatgpt, microsoft bing, google bard and quora poe," *IQ Scores for ChatGPT, Microsoft Bing, Google Bard and Quora Poe (April 7, 2023)*, 2023.

[430] K. M. Caramancion, "News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking," *arXiv preprint arXiv:2306.17176*, 2023.

[431] L. Zhang, "Four tax questions for chatgpt and other language models," 2023.

[432] M. King, "Gpt-4 aligns with the new liberal party, while other large language models refuse to answer political questions,"

[433] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[434] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 2249–2281, 2022.

[435] W. Wang, R. Metzler, and A. G. Cherstvy, "Anomalous diffusion, aging, and nonergodicity of scaled brownian motion with fractional gaussian noise: Overview of related experimental observations and models," *Physical Chemistry Chemical Physics*, vol. 24, no. 31, pp. 18482–18504, 2022.

[436] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2023.

[437] J. Huggins and J. Zou, "Quantifying the accuracy of approximate diffusions and markov chains," in *Artificial Intelligence and Statistics*, pp. 382–391, PMLR, 2017.

[438] W. Feller, "On the theory of stochastic processes, with particular reference to applications, p 403–432," 1949.

[439] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–317, IEEE, 2007.

[440] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," *arXiv preprint arXiv:1710.10686*, 2017.

[441] A. Sedghi, L. J. O'Donnell, T. Kapur, E. Learned-Miller, P. Mousavi, and W. M. Wells III, "Image registration: Maximum likelihood, minimum entropy and deep learning," *Medical image analysis*, vol. 69, p. 101939, 2021.

[442] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your gan: Interactive point-based manipulation on the generative image manifold," in *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

[443] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[444] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[445] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

[446] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

[447] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[448] S. AI, "Stablediffusion2.1 release." https://stability.ai/blog/stablediffusion2-1-release7-dec-2022, 2022. Accessed on August 3, 2023.

[449] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," in *European Conference on Computer Vision*, pp. 105–124, Springer, 2022.

[450] H. Lin, A. Zala, J. Cho, and M. Bansal, "Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning," *arXiv preprint arXiv:2309.15091*, 2023.

[451] A. Munoz, M. Zolfaghari, M. Argus, and T. Brox, "Temporal shift gan for large scale video generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3179–3188, 2021.

[452] M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang, and A. Mallya, "Generative adversarial networks for image and video synthesis: Algorithms and applications," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 839–862, 2021.

[453] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, "Make-a-video: Text-to-video generation without text-video data," *arXiv preprint arXiv:2209.14792*, 2022.

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

[455] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[456] Y. Vasiliev, *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, 2020.

[457] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[458] 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.

[459] E. A. Van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "Chatgpt: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

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

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

[462] 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.

[463] Z. Epstein, A. Hertzmann, I. of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, *et al.*, "Art and the science of generative ai," *Science*, vol. 380, no. 6650, pp. 1110–1111, 2023.

[464] 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.

[465] J. Huang and K. C.-C. Chang, "Citation: A key to building responsible and accountable large language models," *arXiv preprint arXiv:2307.02185*, 2023.

[466] 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.

[467] 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.

[468] F. G. e. a. Baptiste Rozière, Jonas Gehring, "Code llama: Open foundation models for code," *arXiv preprint*, 2023.

[469] T. Calò and L. De Russis, "Leveraging large language models for end-user website generation," in *International Symposium on End User Development*, pp. 52–61, Springer, 2023.

[470] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface," *arXiv preprint arXiv:2303.17580*, 2023.

[471] O. Thawkar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, and F. S. Khan, "Xraygpt: Chest radiographs summarization using medical vision-language models," *arXiv preprint arXiv:2306.07971*, 2023.

[472] Y. Luo, J. Zhang, S. Fan, K. Yang, Y. Wu, M. Qiao, and Z. Nie, "Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine," *arXiv preprint arXiv:2308.09442*, 2023.

[473] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, and X. Gao, "Skingpt-4: An interactive dermatology diagnostic system with visual large language model," *medRxiv*, pp. 2023–06, 2023.

[474] K. Kheiri and H. Karimi, "Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning," *arXiv preprint arXiv:2307.10234*, 2023.

[475] H. Chase, "Langchain, 10 2022," *URL https://github.com/hwchase17/langchain*.

[476] 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.

[477] 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*.

[478] 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.

[479] I. Poola and V. Božić, "These plug-ins revolutionize chatgpt into an everything app-chatgpt's plugin store," 2023.

[480] 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.

[481] H. Gimpel, K. Hall, S. Decker, T. Eymann, L. Lämmermann, A. Mädche, M. Röglinger, C. Ruiner, M. Schoch, M. Schoop, *et al.*, "Unlocking the power of generative ai models and systems such as gpt-4 and chatgpt for higher education: A guide for students and lecturers," tech. rep., Hohenheim Discussion Papers in Business, Economics and Social Sciences, 2023.

[482] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, pp. 1678–1684, 2012.

[483] M. M. Al-Jefri and S. A. Mahmoud, "Context-sensitive arabic spell checker using context words and n-gram language models," in *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, pp. 258–263, IEEE, 2013.

[484] Z. Jiang, J. Araki, H. Ding, and G. Neubig, "How can we know when language models know? on the calibration of language models for question answering," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 962–977, 2021.

[485] L. Wang, W. Zhao, Z. Wei, and J. Liu, "Simkgc: Simple contrastive knowledge graph completion with pre-trained language models," *arXiv preprint arXiv:2203.02167*, 2022.

[486] H. J. Dolfing and I. L. Hetherington, "Incremental language models for speech recognition using finite-state transducers," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.*, pp. 194–197, IEEE, 2001.

[487] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, 2022.

[488] 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.

[489] 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.

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

[491] 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.

[492] 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.

[493] 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.

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

[495] 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.

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

[497] P. Hacker, A. Engel, and M. Mauer, "Regulating chatgpt and other large generative ai models," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1112–1123, 2023.

[498] A. Gokul, "Llms and ai: Understanding its reach and impact," 2023.

[499] 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.

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

[501] 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.

[502] 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.

[503] 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.

[504] 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.

[505] 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.

[506] J. Kaddour, "The minipile challenge for data-efficient language models," *arXiv preprint arXiv:2304.08442*, 2023.

[507] 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.

[508] Z. CHen, L. Cao, S. Madden, J. Fan, N. Tang, Z. Gu, Z. Shang, C. Liu, M. Cafarella, and T. Kraska, "Seed: Simple, efficient, and effective data management via large language models," *arXiv preprint arXiv:2310.00749*, 2023.

[509] J. Yuan, R. Tang, X. Jiang, and X. Hu, "Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability," *arXiv preprint arXiv:2303.16756*, 2023.

[510] Y. Wang, "Deciphering the enigma: A deep dive into understanding and interpreting llm outputs," 2023.

[511] J. V. Lochter, R. M. Silva, and T. A. Almeida, "Deep learning models for representing out-of-vocabulary words," in *Brazilian Conference on Intelligent Systems*, pp. 418–434, Springer, 2020.

[512] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[513] X. Wang, X. Li, Z. Yin, Y. Wu, and L. Jia, "Emotional intelligence of large language models," *arXiv preprint arXiv:2307.09042*, 2023.

[514] S. Doddapaneni, G. Ramesh, M. M. Khapra, A. Kunchukuttan, and P. Kumar, "A primer on pretrained multilingual language models," *arXiv preprint arXiv:2107.00676*, 2021.

[515] D. Schuurmans, "Memory augmented large language models are computationally universal," *arXiv preprint arXiv:2301.04589*, 2023.

[516] Y. Wolf, N. Wies, Y. Levine, and A. Shashua, "Fundamental limitations of alignment in large language models," *arXiv preprint arXiv:2304.11082*, 2023.

[517] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting llm-generated texts," *arXiv preprint arXiv:2303.07205*, 2023.

[518] 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.

[519] 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.

[520] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?," *arXiv preprint arXiv:2305.01937*, 2023.

[521] 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.

[522] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[523] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *arXiv preprint arXiv:2307.10169*, 2023.

[524] T. Fujii, K. Shibata, A. Yamaguchi, T. Morishita, and Y. Sogawa, "How do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study in japanese," *arXiv preprint arXiv:2306.09572*, 2023.

[525] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[526] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[527] H. Laurençon, L. Saulnier, T. Wang, C. Akiki, A. Villanova del Moral, T. Le Scao, L. Von Werra, C. Mou, E. González Ponferrada, H. Nguyen, *et al.*, "The bigscience roots corpus: A 1.6 tb composite multilingual dataset," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31809–31826, 2022.

[528] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*, pp. 2790–2799, PMLR, 2019.

[529] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[530] Y. Chen, S. Qian, H. Tang, X. Lai, Z. Liu, S. Han, and J. Jia, "Longlora: Efficient fine-tuning of long-context large language models," *arXiv preprint arXiv:2309.12307*, 2023.

[531] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[532] M. Pagliardini, D. Paliotta, M. Jaggi, and F. Fleuret, "Faster causal attention over large sequences through sparse flash attention," *arXiv preprint arXiv:2306.01160*, 2023.

[533] S. Liu and Z. Wang, "Ten lessons we have learned in the new" sparseland": A short handbook for sparse neural network researchers," *arXiv preprint arXiv:2302.02596*, 2023.

[534] L. Chen, M. Zaharia, and J. Zou, "Frugalgpt: How to use large language models while reducing cost and improving performance," *arXiv preprint arXiv:2305.05176*, 2023.

[535] S. Chen, S. Wong, L. Chen, and Y. Tian, "Extending context window of large language models via positional interpolation," *arXiv preprint arXiv:2306.15595*, 2023.

[536] R. Li, J. Su, C. Duan, and S. Zheng, "Linear attention mechanism: An efficient attention for semantic segmentation," *arXiv preprint arXiv:2007.14902*, 2020.

[537] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang, "Longt5: Efficient text-to-text transformer for long sequences," *arXiv preprint arXiv:2112.07916*, 2021.

[538] X. Ma, X. Kong, S. Wang, C. Zhou, J. May, H. Ma, and L. Zettlemoyer, "Luna: Linear unified nested attention," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2441–2453, 2021.

[539] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, H. Cao, X. Cheng, M. Chung, M. Grella, K. K. GV, *et al.*, "Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.

[540] T. Dao, D. Y. Fu, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," *arXiv preprint arXiv:2212.14052*, 2022.

[541] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, and N. Zhang, "Editing large language models: Problems, methods, and opportunities," *arXiv preprint arXiv:2305.13172*, 2023.

[542] 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.

[543] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. Ratner, R. Krishna, J. Shen, and C. Zhang, "Large language model as attributed training data generator: A tale of diversity and bias," *arXiv preprint arXiv:2306.15895*, 2023.

[544] N. Gillibrand and C. Draper, "Informational sovereignty: A new framework for ai regulation," *Gillibrand, Nicky, and Chris Draper."Informational Sovereignty: A New Framework For AI Regulation"(July 17, 2023).*, 2023.

[545] D. Oba, M. Kaneko, and D. Bollegala, "In-contextual bias suppression for large language models," *arXiv preprint arXiv:2309.07251*, 2023.

[546] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[547] J. Greene, "Will ChatGPT Make Lawyers Obsolete? (Hint: Be Afraid)," *Reuters*, December 2022.

[548] 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.

[549] 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.

[550] 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.

[551] 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.

[552] 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*.

[553] A. Uchendu, *REVERSE TURING TEST IN THE AGE OF DEEPFAKE TEXTS*. PhD thesis, The Pennsylvania State University, 2023.

[554] 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.

[555] 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*.

[556] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[557] 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.

[558] 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.

[559] 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.

[560] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati, "Large language models still can't plan (a benchmark for llms on planning and reasoning about change)," *arXiv preprint arXiv:2206.10498*, 2022.

[561] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He, "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models," *arXiv preprint arXiv:2303.16421*, 2023.

[562] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, vol. 62, 2022.

[563] M. Hardy, I. Sucholutsky, B. Thompson, and T. Griffiths, "Large language models meet cognitive science: Llms as tools, models, and participants," in *Proceedings of the annual meeting of the cognitive science society*, 2023.

[564] A. Wan, E. Wallace, S. Shen, and D. Klein, "Poisoning language models during instruction tuning," *arXiv preprint arXiv:2305.00944*, 2023.

[565] H. Li, D. Guo, W. Fan, M. Xu, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," *arXiv preprint arXiv:2304.05197*, 2023.

[566] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," *arXiv preprint arXiv:2211.09527*, 2022.

[567] Y. Liu, G. Deng, Y. Li, K. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu, "Prompt injection attack against llm-integrated applications," *arXiv preprint arXiv:2306.05499*, 2023.

[568] C. Zhang, C. Zhang, T. Kang, D. Kim, S.-H. Bae, and I. S. Kweon, "Attack-sam: Towards evaluating adversarial robustness of segment anything model," *arXiv preprint arXiv:2305.00866*, 2023.

[569] L. Chen, M. Zaharia, and J. Zou, "How is chatgpt's behavior changing over time?," *arXiv preprint arXiv:2307.09009*, 2023.

[570] K. Nelson, G. Corbin, M. Anania, M. Kovacs, J. Tobias, and M. Blowers, "Evaluating model drift in machine learning algorithms," in *2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pp. 1–8, IEEE, 2015.

[571] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[572] F. Huang, H. Kwak, and J. An, "Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech," *arXiv preprint arXiv:2302.07736*, 2023.

[573] G. Fergusson, C. Fitzgerald, C. Frascella, M. Iorio, T. McBrien, C. Schroeder, B. Winters, and E. Zhou, "Contributions by,"

[574] P. Li, J. Yang, M. A. Islam, and S. Ren, "Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models," *arXiv preprint arXiv:2304.03271*, 2023.

[575] 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.

[576] S. Biswas, "Potential use of chat gpt in global warming," *Ann Biomed Eng*, vol. 51, pp. 1126–1127, 2023.

[577] 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.

[578] 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.

[579] 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.

[580] 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.

[581] 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.

[582] 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.

[583] A. A. Chien, L. Lin, H. Nguyen, V. Rao, T. Sharma, and R. Wijayawardana, "Reducing the carbon impact of generative ai inference (today and in 2035)," *ACM Hot Carbon 2023*, 2023.

[584] B. Tomlinson, R. W. Black, D. J. Patterson, and A. W. Torrance, "The carbon emissions of writing and illustrating are lower for ai than for humans," *arXiv preprint arXiv:2303.06219*, 2023.

[585] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[586] G. Pistilli, "What lies behind agi: ethical concerns related to llms," *Revue Ethique et Numérique*, 2022.

[587] J. Hill, W. R. Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations," *Computers in human behavior*, vol. 49, pp. 245–250, 2015.

[588] W. Wlasak, S. P. Zwanenburg, C. Paton, *et al.*, "Supporting autonomous motivation for physical activity with chatbots during the covid-19 pandemic: Factorial experiment," *JMIR Formative Research*, vol. 7, no. 1, p. e38500, 2023.

[589] Y. Ma, J. Liu, and F. Yi, "Is this abstract generated by ai? a research for the gap between ai-generated scientific text and human-written scientific text," *arXiv preprint arXiv:2301.10416*, 2023.

[590] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *IFIP international conference on artificial intelligence applications and innovations*, pp. 373–383, Springer, 2020.

[591] M. Mitchell, "How do we know how smart ai systems are?," 2023.

[592] C. Singh, J. X. Morris, J. Aneja, A. M. Rush, and J. Gao, "Explaining patterns in data with language models via interpretable autoprompting," *arXiv preprint arXiv:2210.01848*, 2022.

[593] V. Kumar, A. Choudhary, and E. Cho, "Data augmentation using pre-trained transformer models," *arXiv preprint arXiv:2003.02245*, 2020.

[594] L. Xu, X. Zhang, and Q. Dong, "Cluecorpus2020: A large-scale chinese corpus for pre-training language model," *arXiv preprint arXiv:2003.01355*, 2020.

[595] H. Pan, C. Wang, M. Qiu, Y. Zhang, Y. Li, and J. Huang, "Meta-kd: A meta knowledge distillation framework for language model compression across domains," *arXiv preprint arXiv:2012.01266*, 2020.

[596] M. Ostendorff and G. Rehm, "Efficient language model training through cross-lingual and progressive transfer learning," *arXiv preprint arXiv:2301.09626*, 2023.

[597] S. Iyer, X. V. Lin, R. Pasunuru, T. Mihaylov, D. Simig, P. Yu, K. Shuster, T. Wang, Q. Liu, P. S. Koura, *et al.*, "Opt-iml: Scaling language model instruction meta learning through the lens of generalization," *arXiv preprint arXiv:2212.12017*, 2022.

[598] B. Zhuang, J. Liu, Z. Pan, H. He, Y. Weng, and C. Shen, "A survey on efficient training of transformers," *arXiv preprint arXiv:2302.01107*, 2023.

[599] W. J. Dupps Jr, "Artificial intelligence and academic publishing," *Journal of Cataract & Refractive Surgery*, vol. 49, no. 7, pp. 655–656, 2023.

[600] C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi, "Taking ai risks seriously: a new assessment model for the ai act," *AI & SOCIETY*, pp. 1–5, 2023.

[601] U. M. Fayyad, "From stochastic parrots to intelligent assistants—the secrets of data and human interventions," *IEEE Intelligent Systems*, vol. 38, no. 3, pp. 63–67, 2023.

[602] V. Salikutluk, D. Koert, and F. Jäkel, "Interacting with large language models: A case study on ai-aided brainstorming for guesstimation problems," in *HHAI 2023: Augmenting Human Intellect*, pp. 153–167, IOS Press, 2023.

[603] S. Moore, R. Tong, A. Singh, Z. Liu, X. Hu, Y. Lu, J. Liang, C. Cao, H. Khosravi, P. Denny, *et al.*, "Empowering education with llms-the next-gen interface and content generation," in *International Conference on Artificial Intelligence in Education*, pp. 32–37, Springer, 2023.

[604] K. Nottingham, P. Ammanabrolu, A. Suhr, Y. Choi, H. Hajishirzi, S. Singh, and R. Fox, "Do embodied agents dream of pixelated sheep?: Embodied decision making using language guided world modelling," *arXiv preprint arXiv:2301.12050*, 2023.

[605] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models," *arXiv preprint arXiv:2302.12173*, 2023.

[606] S. Kang, J. Yoon, and S. Yoo, "Large language models are few-shot testers: Exploring llm-based general bug reproduction," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pp. 2312–2323, IEEE, 2023.

[607] B. Goertzel, "Artificial general intelligence: concept, state of the art, and future prospects," *Journal of Artificial General Intelligence*, vol. 5, no. 1, p. 1, 2014.

[608] T. Giannini and J. P. Bowen, "Generative art and computational imagination: Integrating poetry and art," in *Proceedings of EVA London 2023*, pp. 211–219, BCS Learning & Development, 2023.

[609] C. Summerfield, *Natural General Intelligence: How understanding the brain can help us build AI.* Oxford University Press, 2022.

[610] N. Nascimento, P. Alencar, and D. Cowan, "Self-adaptive large language model (llm)-based multiagent systems," *arXiv preprint arXiv:2307.06187*, 2023.

[611] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, *et al.*, "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.

[612] T. Everitt, *Towards safe artificial general intelligence*. PhD thesis, The Australian National University (Australia), 2019.

[613] 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.

[614] C. K. Iyer, F. Hou, H. Wang, Y. Wang, K. Oh, S. Ganguli, and V. Pandey, "Trinity: A no-code ai platform for complex spatial datasets," in *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pp. 33–42, 2021.

[615] B. Allen, S. Agarwal, J. Kalpathy-Cramer, and K. Dreyer, "Democratizing ai," *Journal of the American College of Radiology*, vol. 16, no. 7, pp. 961–963, 2019.

[616] L. Sundberg and J. Holmström, "Democratizing artificial intelligence: How no-code ai can leverage machine learning operations," *Business Horizons*, 2023.

[617] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine learning operations (mlops): Overview, definition, and architecture," *IEEE Access*, 2023.