Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.052618

REVIEW

Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models

Zheyi Chen[1,#], Liuchang Xu[1,#], Hongting Zheng[1], Luyao Chen[1], Amr Tolba[2,3], Liang Zhao[4], Keping Yu[5,*] and Hailin Feng[1,*]

[1]College of Mathematics and Computer Science, Zhejiang A&F University, Hangzhou, 311300, China

[2]Computer Science Department, Community College, King Saud University, Riyadh, 11437, Saudi Arabia

[3]Mathematics and Computer Science Department, Faculty of Science, Menofia University, Shebin El Kom, Menoufia Governorate, 32511, Egypt

[4]School of Computer Science, Shenyang Aerospace University, Shenyang, 110136, China

[5]Graduate School of Science and Engineering, Hosei University, Tokyo, 184-8584, Japan

*Corresponding Authors: Keping Yu. Email: keping.yu@ieee.org; Hailin Feng. Email: hlfeng@zafu.edu.cn

#Zheyi Chen and Liuchang Xu contributed equally to this work

Received: 09 April 2024 Accepted: 09 July 2024 Published: 15 August 2024

ABSTRACT

Since the 1950s, when the Turing Test was introduced, there has been notable progress in machine language intelligence. Language modeling, crucial for AI development, has evolved from statistical to neural models over the last two decades. Recently, transformer-based Pre-trained Language Models (PLM) have excelled in Natural Language Processing (NLP) tasks by leveraging large-scale training corpora. Increasing the scale of these models enhances performance significantly, introducing abilities like context learning that smaller models lack. The advancement in Large Language Models, exemplified by the development of ChatGPT, has made significant impacts both academically and industrially, capturing widespread societal interest. This survey provides an overview of the development and prospects from Large Language Models (LLM) to Large Multimodal Models (LMM). It first discusses the contributions and technological advancements of LLMs in the field of natural language processing, especially in text generation and language understanding. Then, it turns to the discussion of LMMs, which integrates various data modalities such as text, images, and sound, demonstrating advanced capabilities in understanding and generating cross-modal content, paving new pathways for the adaptability and flexibility of AI systems. Finally, the survey highlights the prospects of LMMs in terms of technological development and application potential, while also pointing out challenges in data integration, cross-modal understanding accuracy, providing a comprehensive perspective on the latest developments in this field.

KEYWORDS

Artificial intelligence; large language models; large multimodal models; foundation models

1 Introduction

Language serves as a foundational element in human communication and expression, as well as in the interaction between humans and machines, necessitating the development of generalized models to empower machines with the ability to perform complex linguistic tasks. The need for generalized models stems from the growing demand for machines to handle complex language tasks, including translation, summarization, information retrieval, conversational interactions [1], etc. This necessity is rooted in the intrinsic human capability to communicate and express thoughts. Language is a prominent ability in human beings to express and communicate, which develops in early childhood and evolves over a lifetime [2,3]. Unlike humans, machines lack the innate ability to comprehend and generate human language, a gap that can only be bridged through the deployment of sophisticated artificial intelligence (AI) algorithms. It has been a longstanding research challenge to achieve this goal, to enable machines to read, write, and communicate like human [4]. Addressing this challenge, the field of language modeling aims to advance machine language intelligence by focusing on the generative likelihood of sequences of words, thereby enabling the prediction of future or missing tokens [5]. This pursuit has been a focal point of research, evolving through four significant stages, each marking a progressive step towards enabling machines to read, write, and communicate with human-like proficiency.

Building upon these four stages, the development of LMMs emerges as a pivotal fifth stage in the evolution of artificial intelligence.

LMMs mark a significant advancement in AI by integrating multisensory skills like visual understanding and auditory processing with the linguistic capabilities of LLMs. This approach not only leverages the dominant role of vision but also emphasizes the importance of other modalities such as sound, enhancing AI systems to be more adept and versatile. By incorporating a broader range of sensory inputs, LMMs aim to achieve a more powerful form of general intelligence, capable of efficiently performing a wider array of tasks. The five developmental stages are detailed as follows.

1. *Statistical Language Models (SLM)*

SLMs are a type of language model that uses statistical methods to predict the probability of a sequence of words in a language. These models are based on the assumption that the likelihood of a word occurring in a text depends on the words that precede it. SLMs [6–9] analyze large corpora of text to learn word occurrence patterns and relationships. SLMs have been fundamental in various NLP [10] like speech recognition, text prediction, and machine translation before the rise of more advanced neural network-based models. Therefore, the specially proposed principles and methods [11] are used to alleviate the problems encountered in information retrieval challenges.

2. *Neural Language Models (NLM)*

NLMs are a type of language model that uses neural networks, especially deep learning techniques, to understand and generate human language. Unlike SLMs that rely on counts and probabilities of sequences of words, neural models use layers of artificial neurons to process and learn from large amounts of text data. These models capture complex patterns and dependencies in language, allowing for more accurate and contextually relevant language generation and understanding. Examples include Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), and transformer models like Generative Pre-trained Transformer (GPT). They are widely used in applications like machine translation, text generation, and speech recognition.

3. Pre-Trained Language Models (PLM)

PLMs are a category of language model that have been previously trained on large datasets before being used for specific tasks. This pre-training involves learning from vast amounts of text data to understand the structure, nuances, and complexities of a language. Once pre-trained, these models can be fine-tuned with additional data specific to a particular application or task, such as text classification, question answering, or language translation. Examples of PLMs include Bidirectional Encoder Representations from Transformers (BERT) [12], GPT, and Text-to-Text Transfer Transformer (T5) [13]. These models have revolutionized the field of NLP by providing a strong foundational understanding of language, which can be adapted to a wide range of language-related tasks.

4. Large Language Models (LLM)

Artificial Intelligence, particularly generative AI, has garnered widespread attention for its capacity to produce lifelike outputs [14]. LLMs are complex computational systems in the field of artificial intelligence, particularly NLP. They are typically constructed using deep learning techniques, often leveraging Transformer architectures. These models are characterized by their vast number of parameters, often in the billions, which enable them to capture a wide range of linguistic nuances and contextual variations. LLMs are trained on extensive corpora of text, allowing them to generate, comprehend, and interact using human language with a high degree of proficiency. Their capabilities include but are not limited to text generation, language translation, summarization, and question-answering. These models have significantly advanced the frontiers of NLP, offering more sophisticated and context-aware language applications.

5. Large Multimodal Models (LMM)

LMMs are advanced artificial intelligence systems capable of processing and understanding multiple types of data inputs. In addition to standard text-based applications, LLMs are expanding their capabilities to engage with various forms of media, such as images [15,16], videos [17,18] and audio files [19,20] among others. They are multimodal because they can integrate and interpret information from these varied modes simultaneously. These models leverage large-scale datasets and sophisticated neural network architectures to learn complex patterns across different data types. This ability allows them to perform tasks like image captioning, where they generate descriptive text for images, or answer questions based on a combination of text and visual information.

The emergence of LLMs has revolutionized our ability to process and generate human-like text, thereby enhancing applications in numerous fields such as automated customer support, content creation, and language translation, and opening up new possibilities for human-computer interaction. As we advance beyond the realm of pure text-based interactions, the field is witnessing the rise of LMMs. These sophisticated models are pioneering the integration of multisensory data, notably visual and auditory inputs, to better emulate the comprehensive sensory experiences that are central to human cognition. In the field of computer vision, efforts are being made to create vision-language models akin to ChatGPT, aiming to enhance multimodal dialogue capabilities [21–24]. GPT-4 [25] has already taken strides in this direction by accommodating multimodal inputs and incorporating visual data. This progression can be charted through the development of attention mechanisms in LLMs, which have been instrumental in improving contextual understanding. Attention has evolved from a fundamental concept to more complex types and variations, each with its own optimization challenges. Building upon this, the architecture of LMMs incorporates these advanced attention frameworks to process and synthesize information across multiple modalities. Concurrently, the field grapples with open issues such as contextual understanding inherent limitations, the challenges of ambiguity and vagueness in language, and the phenomenon of catastrophic forgetting. Foundational LLMs sometimes misinterpret instructions and "hallucinate" facts, undermining their practical effectiveness [26]. Therefore, there is a focus on correcting hallucinations and enhancing cognitive abilities within these models, which is critical to their reliability and effectiveness. Addressing these challenges requires innovative training data and methods tailored to LLMs and LMMs alike, aiming to refine internal and external reasoning processes. This fine-tuning is essential for achieving accuracy in reasoning, enabling these models to make informed decisions based on a combination of learned knowledge and real-time sensory input. The practical applications of LLMs and LMMs are expansive and transformative, particularly in sectors like healthcare, where they can interpret patient data to inform diagnoses and treatments, and in finance, where they can analyze market trends for more accurate forecasting. In robotics, LMMs facilitate more natural human-robot interactions and enable machines to navigate and interact with external environments more effectively. In sum, the journey from LLMs to LMMs is not merely an incremental step but a significant leap towards creating AI systems that can understand and interact with the world in a manner akin to human intelligence. The eventual integration of these models into real-world applications promises to enhance the efficacy and sophistication of AI technologies across the board.

This review and its subsequent exposition aim to detail the current landscape and future direction of LLMs and LMMs, exploring the nuanced details of these models and their transformative potential in multimodal AI. The structure and content of the article are as shown in Figs. 1 and 2. Fig. 1 provides a broad overview of LLMs and LMMs in six areas: 1. Attention Mechanism 2. Structure 3. Training Methods 4. Training Data 5. Open Issues 6. Applications. Through the Sankey diagram, Fig. 2 counts 325 documents, including 45 in proceedings, 72 other articles, 205 articles, and 3 books. Fig. 3 illustrates the timeline of model proposals from 2019 to mid-2023, with dark blue indicating multi-modal models. The pie chart depicts the proportion of multimodal and non-multimodal models from 2021 to 2023. It is evident from the picture that the development and application of multimodal models are becoming increasingly recognized and embraced by the public.

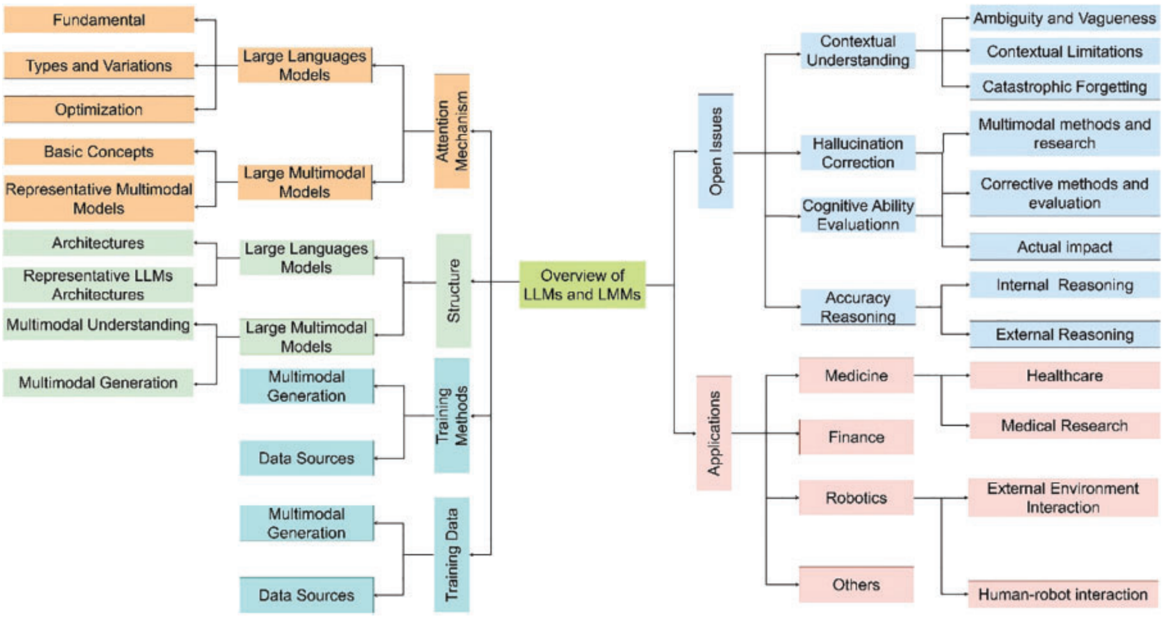

Figure 1: Overview of LLMs and LMMs

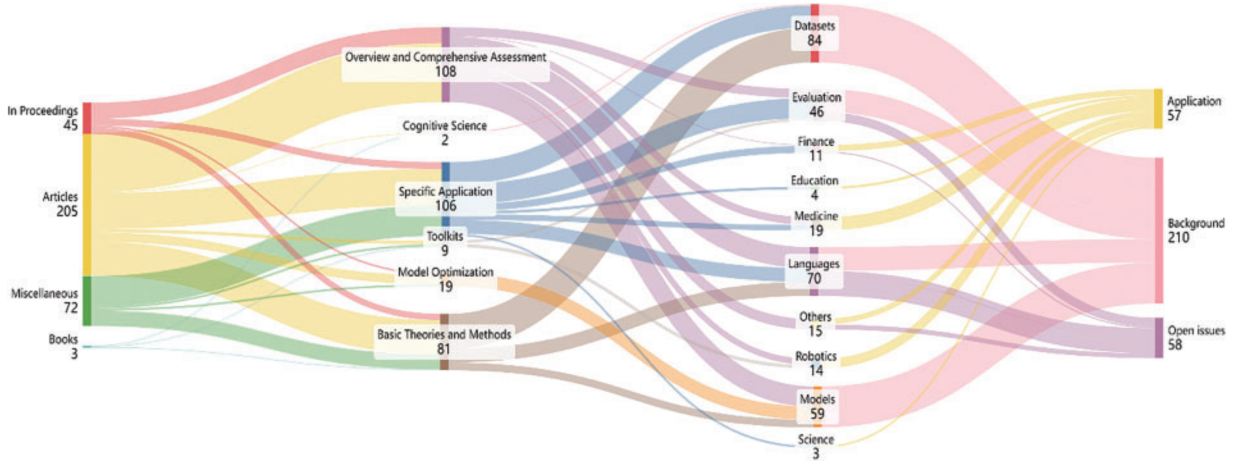

Figure 2: Literature source sankey diagram

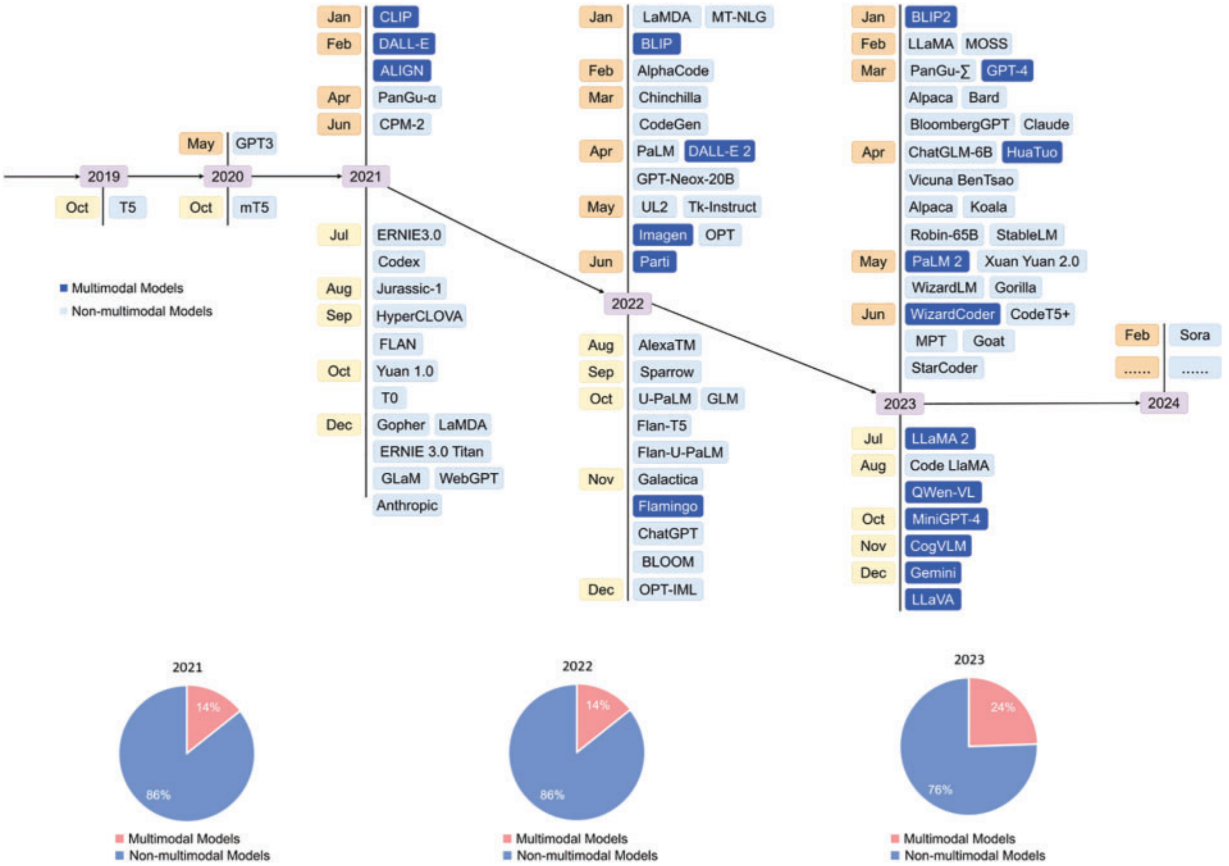

Figure 3: Multimodal model growth from 2019 to 2024

Adhering to the hierarchical structure of the outline, with a focus on large language models and multimodal models as primary keywords, we meticulously curated a collection of representative documents. These documents, numbering approximately 325 in total, span multiple fields and were selected based on criteria such as research background, unresolved issues, and practical applications.

The main contributions of this paper are as follows:

- We offer an examination of the evolution and future potential spanning from LLMs to LMMs. Initially, we delve into the contributions and technological strides of LLMs within the realm of NLP, particularly in the arenas of text generation and linguistic comprehension.
- Subsequently, we transition to an exploration of LMMs, amalgamating diverse data modalities such as textual, visual, and auditory inputs, thereby show-casing sophisticated proficiencies in cross-modal comprehension and content generation. This pioneering integration opens avenues for enhanced adaptability and versatility within AI systems.
- We have elucidated the prospects of LMMs in terms of technological advancement and application potential, while also revealing the main challenges related to data integration, cross-modal understanding accuracy, and the current landscape. Finally, in specific application domains, we not only detailed the differences between LLMs and LMMs but also highlighted the necessity of their transformation in concrete applications.

2 Background

LLMs, like GPT-3 [27], PaLM [28], Galactica [29], and LLaMA [30], are transformer-based models with hundreds of billions of parameters, trained on vast text datasets [31]. These models are adept at comprehending natural language and executing complex tasks through text generation. This section provides an overview of LLMs, covering attention mechanisms, model architecture, and training data and methods for a concise understanding of their operation.

2.1 Attention in LLMs

Attention in LLMs is fundamental for processing and understanding complex language structures. It operates by focusing on specific parts of the input data, thereby discerning relevant context and relationships within the text. Various types and variations of attention exist, such as self-attention and multi-head attention, each offering unique advantages in handling different understanding. Optimization of attention mechanisms, especially in large models, involves techniques like sparse attention to manage computational efficiency and memory usage, crucial for scaling LLMs for more extensive and intricate datasets. These mechanisms collectively enhance the LLMs ability to generate coherent, contextually relevant responses, making them versatile in numerous language processing applications.

We will delve into the concept of Attention in LLMs from three distinct perspectives: the fundamental principles underlying attention mechanisms, the types and variations that exist within these models, and the optimization strategies employed to enhance their efficiency and effectiveness.

2.1.1 Fundamental

The attention mechanism plays a pivotal role within the Transformer framework, facilitating inter-token interactions across sequences and deriving representations for both input and output sequences. When processing sequential data, the attention mechanism simulates the human attention process by endowing the model with varying degrees of attention to different parts. This is akin to how humans allocate attention when processing information, allowing the model to selectively focus on specific parts of the sequence, rather than treating all information equally. In the attention mechanism, the model calculates weights based on different parts of the input data to indicate which parts the model should pay more attention to. These weights determine which positions in the input sequence should be considered when calculating the output, enabling the model to simulate selective attention similar to that of humans when processing sequential data.

In summary, the attention mechanism simulates the human attention process by dynamically adjusting attention to different parts, enabling the model to better comprehend and process sequential data. This contributes to the improvement of performance in models for tasks such as NLP and machine translation.

2.1.2 Types and Variations

1. *Self-Attention*

Self-attention [32] within the Transformer model is characterized as a mechanism that calculates a position response in a sequence by considering all positions, assigning importance through learned patterns. Unlike models that process sequences sequentially, this allows for parallel processing, enhancing training efficiency. It adeptly handles long-range dependencies, crucial in translation where contextual comprehension is key. By evaluating interrelations across an entire input sequence, self-attention significantly improves the model ability to interpret intricate patterns and dependencies, a distinct advantage in complex task handling. Self-attention parallel processing capability and proficiency in capturing long-range dependencies offer a significant advantage over traditional models like RNN and convolutional neural networks (CNN). Particularly in NLP, where understanding long-distance relationships in data is crucial, self-attention proves more effective. This mechanism is extensively utilized in the Transformer model, now a dominant framework in various NLP tasks, including machine translation, text generation, and classification. The transformer reliance on self-attention, avoiding RNN or CNN architectures, markedly enhances its ability to handle extended sequences, driving breakthroughs in performance across multiple applications.

2. *Cross Attention*

Cross-attention, a pivotal concept in deep learning and NLP, denotes an advanced attention mechanism. This mechanism is instrumental in enabling models to intricately associate and assign weights to elements across two distinct sequences. A prime example is in machine translation, where it bridges the source and target language components. Essentially, cross-attention empowers a model to integrate and process information from one sequence while attentively considering another sequence context. Predominantly utilized in sequence-to-sequence models, which are prevalent in machine translation, text summarization, and question-answering systems, cross-attention plays a crucial role. It facilitates the model ability to decode and interrelate the intricacies between an input sequence (like a segment of text) and its corresponding output sequence (such as a translated phrase or an answer). Delving into specifics, the cross-attention mechanism leverages insights from one sequence (for instance, the output from an encoder) to attentively navigate and emphasize particular segments of another sequence (like input to a decoder). This capability is key to grasping and handling the multifaceted dependencies existing between sequences, thereby significantly boosting precision and efficiency of the model in complex tasks. In the realm of machine translation, for instance, cross-attention enables the model to meticulously concentrate on specific fragments of the source language sentence, ensuring a more accurate and contextual translation into the target language. Overall, cross-attention stands as a cornerstone technique in deep learning, essential for the nuanced understanding and processing of complex inter-sequential relationships.

3. Full Attention

Full attention in the context of neural network architectures, particularly Transformers, refers to a mechanism where each element in a sequence attends to every other element. Unlike sparse attention which selectively focuses on certain parts of the sequence, full attention involves calculating attention scores between all pairs of elements in the input sequence. This approach, while computationally intensive due to its quadratic complexity, provides a comprehensive understanding of the relationships within the data, making it highly effective for tasks requiring deep contextual understanding. However, its computational cost limits its scalability, especially for very long sequences.

4. Sparse Attention

Sparse attention, as detailed in the work [33], revolutionizes the efficiency of Transformer models for extensive sequences. By introducing sparse factorizations into the attention matrix, this approach transforms the computational complexity from quadratic to nearly linear. This reduction is accomplished through a strategic decomposition of the full attention mechanism into more manageable operations that closely emulate dense attention, but with significantly reduced computational demands. Demonstrating its robustness, the sparse Transformer excels in processing a wide range of data types, including text, images, and audio, thereby setting new performance benchmarks in density modeling for complex datasets like Enwik8 and CIFAR-10. Notably, its design allows for the handling of sequences up to a million steps in length, a feat that significantly surpasses the capabilities of standard Transformer models. This advancement not only enhances the efficiency of processing lengthy sequences but also opens new avenues for complex sequence modeling tasks.

5. Multi-Query/Grouped-Query Attention

Multi-query attention is a variant of the traditional attention mechanism commonly used in neural network architectures. In this approach, instead of computing attention with a single query per input element, multiple queries are used simultaneously for each element [34]. This allows the model to capture a wider range of relationships and interactions within the data. Multi-query attention can provide a richer and more nuanced understanding of the input, as it enables the model to attend to different aspects or features of the data in parallel, enhancing its ability to learn complex patterns and dependencies.

6. Flash Attention

Flash attention [35] is a technique in neural network architecture that optimizes the efficiency of attention mechanisms, specifically in Transformers. It focuses on improving the speed and reducing memory usage during the attention calculation process. This is achieved by enhancing the handling of memory read and write operations, particularly in GPUs. As a result, flash attention can operate significantly faster than traditional attention mechanisms, while also being more memory efficient. This makes it particularly advantageous for tasks involving large-scale data processing or when operating under memory constraints. It employs a tiling strategy to minimize memory transactions between different GPU memory levels, leading to faster and more memory-efficient exact attention computation. Notably, flash attention design allows it to perform computations up to several times faster than traditional attention methods, with substantial memory savings, marking a significant advancement in handling large-scale data in neural networks. FlashAttention has been implemented as a fused kernel within CUDA and is now integrated into prominent frameworks including PyTorch [36], DeepSpeed [37], and Megatron-LM [38]. This integration signifies its practical applicability and enhancement of these platforms, providing a more efficient and memory-effective approach to attention computation in large-scale neural network models.

FlashAttention-2 [39], as an evolution of the original FlashAttention, significantly advances the efficiency of attention processing in Transformers. It optimizes GPU utilization by refining the algorithm for more effective work partitioning. The enhancement involves minimizing non-matrix multiplication operations and enabling parallel processing of attention, even for individual heads, across various GPU thread blocks. Moreover, it introduces a more balanced distribution of computational tasks within thread blocks. These strategic improvements result in FlashAttention-2 achieving approximately double the speed of its predecessor, nearing the efficiency of optimized matrix multiplication operations, thereby marking a substantial leap in executing large-scale Transformer models efficiently.

7. PagedAttention

PagedAttention is an advanced technique in neural net-work architecture, specifically designed to address the limitations of conventional attention mechanisms in handling long sequences. It operates by dividing the computation into smaller, more manageable segments or pages, thereby reducing the memory footprint and computational load. This approach allows for efficient processing of long sequences that would otherwise be challenging or infeasible with standard attention models, making it particularly useful in large-scale NLP and other data-intensive applications. The PagedAttention method has been developed to optimize the use of memory and augment the processing capacity of LLMs in operational environments [40].

2.1.3 Optimization

In the realm of neural network models, particularly in NLP, attention mechanisms have emerged as a pivotal innovation, enhancing model accuracy and contextual understanding. These mechanisms enable models to selectively concentrate on relevant segments of input data, effectively capturing intricate dependencies and relations. This targeted focus significantly bolsters performance in tasks such as language translation and summarization.

However, with the growing scale of models and data, traditional attention mechanisms often grapple with increased computational demands, affecting training speed and efficiency. To address these challenges, optimization techniques like sparse attention have been developed. Sparse attention streamlines the process by selectively focusing on crucial data points, thus reducing the computational burden. This selective approach not only expedites the training process but also scales more adeptly with larger datasets.

The incorporation of sparse attention is crucial in managing the escalating complexities and sizes of datasets in advanced deep learning applications. By optimizing attention mechanisms, we can build models that are not only more accurate but also faster and more scalable, catering to the demanding requirements of modern machine learning tasks. These advancements are integral to pushing the boundaries of what neural network models can achieve, particularly in processing and understanding large-scale, complex data structures.

2.2 Attention in LMMs

2.2.1 Basic Concepts

Multimodal learning in computer science involves integrating data from various modalities, such as text, images, and sound, to create models that understand and process information more holistically.

This approach is crucial for tasks requiring an understanding across multiple data types. For instance, in a scenario where both visual cues from images and descriptive cues from text are essential, multimodal learning enables the model to combine these distinct types of information to generate a more accurate and comprehensive understanding. This is particularly important in complex data interpretation and decision-making tasks, where relying on a single modality might lead to incomplete or biased conclusions. Multimodal learning, therefore, plays a vital role in enhancing the depth and breadth of data analysis and interpretation in AI applications.

In contemporary multimodal research, image-text conversion and matching represent a pivotal area, involving the precise alignment of visual content with textual descriptions. Attention mechanisms play an essential role in this process. By incorporating cross-modal attention mechanisms, LMMs are able to focus on parts of the image that are closely related to the text descriptions, thereby achieving more accurate local alignments. Additionally, these models utilize global attention to integrate the semantic information of the entire image and text, ensuring overall semantic consistency. Xu et al. [41] proposed a novel multimodal model named Cross-modal Attention with Semantic Consistency (CASC). It employs an innovative attention mechanism to integrate both local and global matching strategies. Through finely-tuned cross-modal attention, it achieves granular local alignment, while the use of multi-label prediction ensures the consistency of global semantics. This strategy of combining local and global perspectives through attention not only enhances the accuracy of image-text matching but also significantly improves the model's ability to handle complex multimodal information.

Cai et al. [42] proposed a graph-attention based multimodal fusion network for enhancing the joint classification of hyperspectral images and LiDAR data. It includes an HSI-LiDAR feature extractor, a graph-attention fusion module, and a classification module. The fusion module constructs an undirected weighted graph with modality-specific tokens to address long-distance dependencies and explore deep semantic relationships, which are then classified by two fully connected layers.

In modern AI research, multimodal models integrate diverse data like text, images, and sound using cross-modal attention mechanisms. These mechanisms allow models to focus on relevant information across modalities. For example, in image-text matching tasks, they enable identification of key text elements and their alignment with corresponding visual details. This enhances data processing accuracy and improves the model's adaptability and problem-solving capabilities in complex scenarios, proving essential for multimodal tasks.

2.2.2 Representative Multimodal Models

This section offers a comprehensive exploration of how advanced AI models integrate varied data types like text, images, and audio. These multimodal models, adept at processing and interpreting multi-sensory information, facilitate a nuanced understanding of complex data sets. The introduction outlines the architecture of these models, their data fusion techniques, and diverse applications in fields such as NLP, computer vision, and human-computer interaction, setting the groundwork for appreciating their interdisciplinary and technological complexity. It has been shown that CLIP-NAV [43] explores an innovative approach to Vision-and-Language Navigation (VLN) using the CLIP model for zero-shot navigation. The focus is on improving VLN in diverse and previously unseen environments without dataset-specific fine-tuning. The study leverages CLIP strengths in language grounding and object recognition to navigate based on natural language instructions. The results demonstrate that this approach can surpass existing supervised baselines in navigation tasks, highlighting the potential of CLIP in generalizing better across different environments for VLN tasks. This research marks a significant stride in the field of autonomous navigation using language models.

In addition, RoboCLIP [44] addresses the challenge of efficiently teaching robots to perform tasks with minimal demonstrations. The method utilizes a single demonstration, which can be a video or textual description, to generate rewards for online reinforcement learning. This approach eliminates the need for labor-intensive reward function designs and allows for the use of demonstrations from different domains, such as human videos. RoboCLIP use of pretrained Video-and-Language Models (VLMs) without fine-tuning represents a significant advancement in enabling robots to learn tasks effectively with limited data. Table 1 provides an extensive overview of frequently used non-multimodal and multimodal models, delineating their distinct characteristics and primary functions. It utilizes a color-coding system where non-multimodal models are differentiated as follows: transformer-based models are highlighted in yellow, code generation models in pink, and multilingual or cross-language models in gray. This color coding helps to clearly understand each model type and its functionalities.

Table 1: Comparison of characteristics and principal tasks of multimodal and non-multimodal models

Models	Insights
T5 [13]	The unified framework simplifies model design and training processes.
GPT-3 [27]	Revolutionary AI for comprehensive, human-like text generation across various domains.
CC595K [45]	CC595K is used alongside WebVid2M for initial vision branch training to help the model understand audio.
ERNIE3.0 [46]	Advanced language model emphasizing contextual understanding, supporting diverse NLP tasks.
PanGu-α [47]	It focuses on extensive pretraining for enhanced natural language understanding and generation.
CPM-2 [48]	Large-scale pretraining enhances understanding and generation of natural language text.
ERNIE 3.0 Titan [49]	High-performing language model for advanced contextual understanding and NLP.
GPT-NepX-20B [50]	Advanced language model optimized for comprehensive text generation across diverse tasks and domains.
BLOOM [51]	Innovative algorithm for efficient and scalable data filtering and retrieval.
GLaM [52]	Pretraining corpus choice greatly affects LLMs' performance in downstream tasks.
LaMDA [53]	Refining various external information through fine-tuning the model.
UL2 [54]	Enhanced performance on downstream tasks is facilitated by mode switching training.
GLM-130B [55]	Model performance is boosted by employing pretraining data.
Jurassic-1 [56]	High-performance language model prioritizing efficiency and accuracy for diverse NLP tasks.
HyperCLOVA [57]	A cutting-edge AI system designed for versatile and efficient natural NLP tasks.
Yuan 1.0 [58]	A state-of-the-art language model engineered for robust and versatile natural language understanding and generation tasks.
PanGu-Σ [59]	Sparse models are characterized by lower computational costs.

(Continued)

Table 1 (continued)

Models	Insights
XuanYuan 2.0 [60]	In order to improve the memory ability of the model, pre-training and fine-tuning will be combined in the training.
OPT [61]	Optimization process for maximizing efficiency and performance in various computational tasks.
Gopher [62]	Efficient, typed language for simple, productive web development.
Galactica [29]	Galactica's performance improves consistently across different benchmarks, exceeding previous LLMs research.
Chinchilla [63]	Scaling model size and training token count proportionally yields optimal computation.
CoQA [64]	The primary goal is to challenge and improve conversational AI systems with realistic, complex question-answering scenarios.
LLaMA [30]	Model performance is improved through scaling.
PaLM [28]	Larger models tend to have better memory capacity during training.
U-PaLM [65]	Through training with a mixed denoiser, the filling capability and diversity of open-text generation have been enhanced.
AlexaTM [66]	Adding auxiliary tasks can enhance the model's contextual learning efficiency.
BloombergGPT [67]	Combining general and specialized data enhances model performance without limiting capabilities.
mT5 [68]	Multitask T5, excels in multilingual tasks, showcasing versatile language understanding.
AlphaCode [69]	Utilizing encoder and decoder to present an asymmetric transformer model, thereby enhancing efficiency.
CodeGen [70]	Utilizing distributed prompts to generate code and synthesize will better understand user intent.
CodeT5+ [71]	Set multiple training goals for better performance.

In contrast, Table 2 focuses exclusively on multimodal models. It categorizes these models based on their capabilities with a similar color-coding system: models capable of generating images from textual descriptions and understanding the relationship between images and text are marked with yellow; those that demonstrate a comprehensive grasp of multimodal data are in pink; and models that offer a holistic approach to both multimodal comprehension and generative abilities are designated with gray.

Table 2: Overview of understanding and generation capabilities in multimodal models

Models	Insights
DALL-E [72]	Construct new scenes by transferring memory information.
DALLE-E2 [73]	CLIP embeddings utilized for diverse, photorealistic image generation while preserving semantic and stylistic integrity.

(Continued)

Table 2 (continued)

Models	Insights
Imagen [74]	Leveraging large language models for unprecedented photorealism and precise text-image alignment.
Parti [75]	Generates photorealistic images from text, leveraging Transformer-based encoding.
MiniGPT-4 [76]	MiniGPT-4 aligns visual and language models, revealing advanced multi-modal abilities.
ALIGN [77]	Utilizing extensive noisy image-text data for cutting-edge visual and language representations.
CLIP [78]	Using a universal visual encoder, zero-sample visual recognition can be achieved.
BLIP2 [79]	Superior performance, fewer parameters, enabling zero-shot image-to-text guided generation.
COGVLM [80]	A deep fusion model achieving state-of-the-art performance on cross-modal benchmarks.
Flamingo [15]	Versatile visual language models adept at rapid adaptation to novel tasks with minimal annotated examples.
GPT4 [25]	A transformer model achieving human-level performance on various benchmarks.
PaLM-2 [81]	Enhanced multilingual Transformer: faster inference, stable performance.
WizardCoder [82]	Levates Code LLMs with fine-tuned instruction, outperforming all Open-source and closed LLMs on code tasks.
LLaMA2 [30]	A set of LLMs fine-tuned for dialogue, surpassing open-source models in benchmarks and safety evaluations.
QWen-VL [83]	Advanced LVLMs, setting new benchmarks in vision-language understanding.
Moe-llava [84]	End-to-end training connects visual encoder with LLMs for universal understanding.

2.3 Structure of LLMs

2.3.1 Architectures

In this discussion, we explore the diverse variants of Transformer architectures, which stem from variations in how attention mechanisms are applied and how transformer blocks are interconnected.

1. *Encoder-Only*

In the landscape of NLP, the advent of the encoder-only architecture signifies a pivotal development. This architecture departs from traditional sequential processing by employing the Transformer encoder to generate rich, contextual representations of input text. The innovation lies in its ability to grasp the nuances of language through bidirectional context, making it adept across a spectrum of NLP tasks, including but not limited to text classification, entity recognition, and sentiment analysis. The encoder-only model capacity for pre-training on extensive corpora before fine-tuning for specific tasks has set a new standard for understanding and processing human language, thus reshaping the methodologies employed in NLP research and applications.

2. *Encoder-Decoder*

The encoder-decoder architecture, a cornerstone in the field of NLP, has been instrumental in advancing machine translation, text summarization, and question answering systems. This framework, as detailed by Sutskever et al. [85], employs a dual-component approach where the encoder processes the input sequence to a fixed-length vector representation, which the decoder then uses to generate the target sequence. This separation allows the model to handle variable-length inputs and outputs, enabling a more flexible and accurate translation of complex language structures. The architecture efficacy in capturing long-distance dependencies and its adaptability to various sequential tasks have catalyzed significant innovations in NLP, making it a fundamental model for researchers and practitioners alike.

3. *Decoder-Only*

The decoder-only architecture, notably pioneered by models such as GPT [86], represents a transformative approach in the realm of NLP, particularly in text generation tasks. This architecture, eschewing the encoder component, focuses solely on the decoder to predict the next token in a sequence based on the preceding ones. It leverages an autoregressive model that processes text in a sequential manner, ensuring that each prediction is contingent upon the tokens that came before it. This design facilitates the generation of coherent and contextually relevant text, making decoder-only models particularly adept at tasks such as story generation, creative writing, and more. The effectiveness of this architecture has been demonstrated across various domains, showcasing its versatility and power in capturing the nuances of human language.

Causal Decoder: The primary goal of a LLMs is to forecast the subsequent token given the preceding sequence of tokens. Although incorporating additional context from an encoder can enhance the relevance of predictions, empirical evidence suggests that LLMs can still excel without an encoder [87], relying solely on a decoder. This approach mirrors the decoder component of the traditional encoder decoder architecture, where the flow of information is unidirectional, meaning the prediction of any token t_k is contingent upon the sequence of tokens leading up to and including t_{k-1}. This decoder-only configuration has become the most prevalent variant among cutting-edge LLMs.

Prefix Decoder: In encoder-decoder architectures, causal masked attention allows the encoder to utilize self-attention to consider every token within a sentence, enabling it to access tokens from t_{k+1} to t_n as well as those from t_1 to t_{k-1} when computing the representation for t_k. However, omitting the encoder in favor of a decoder-only model removes this comprehensive attention capability. A modification in decoder-only setups involves altering the masking strategy from strictly causal to permitting full visibility for certain segments of the input sequence, thereby adjusting the scope of attention and potentially enhancing model flexibility and understanding [1].

2.3.2 *Representative LLMs Architectures*

In the evolving landscape of NLP, various model architectures have significantly advanced the field. The T5 [13] and BART [88] models, both embodying the encoder-decoder architecture, have redefined versatility in NLP tasks. T5 converts all NLP problems into a unified text-to-text format, leveraging a comprehensive Transformer architecture for both encoding and decoding phases. Similarly, BART integrates the bidirectional encoding capabilities of BERT with the autoregressive decoding prowess of GPT, enhancing performance across a range of generative and comprehension tasks. On the other hand, the encoder-only architecture is exemplified by models such as RoBERTa [89] and ALBERT [90]. RoBERTa refines the BERT framework through optimized pre-training techniques, achieving superior results on benchmark tasks. ALBERT reduces model size and increases training speed without compromising performance, illustrating the efficiency of architecture optimization. XLNet [91] and TransformerXL [92] explore the realm of Causal Decoder architecture. XLNet integrates the best of Transformer self-attention with autoregressive language modeling, capturing bidirectional context in text sequences through permutation language modeling. Transformer-XL introduces a novel recurrence mechanism to handle longer text sequences, effectively capturing long-range dependencies. Besides, ELECTRA [93] represents a unique take on the encoder-only architecture, introducing a novel pre-training task that distinguishes between real and artificially replaced tokens to train the model more efficiently. These architectures, through their innovative designs and applications, underscore the dynamic and rapidly advancing nature of machine learning research in NLP, each contributing unique insights and capabilities to the domain.

2.4 Structure of LMMs

2.4.1 Multimodal Understanding

1. Modality encoder

Modality encoder is tasked with encoding inputs from diverse modalities to obtain corresponding features [94]. Fig. 4 depicts iconic models and notable representatives at various junctures in time, showcasing the evolving landscape of tasks undertaken by these models over different epochs. As science and technology progress, the pervasive adoption of large language models and multi-modal languages across diverse domains is poised to proliferate, facilitating the execution of a myriad of distinct tasks.

Visual Modality: For image processing, four notable encoders are often considered. NFNet-F6 [95] represents a modern take on the traditional ResNet architecture, eliminating the need for normalization layers. It introduces an adaptive gradient clipping technique that enhances training on highly augmented datasets, achieving state-of-the-art (SOTA) results in image recognition. Vision Transformer (ViT) [96] brings the Transformer architecture, originally designed for NLP, to the realm of images. By dividing images into patches and applying linear projections, ViT processes these through multiple Transformer blocks, enabling deep understanding of visual content. CLIP ViT [78] bridges the gap between textual and visual data. It pairs a vision Transformer with a text encoder, leveraging contrastive learning from a vast corpus of text-image pairs. This approach significantly enhances the model's ability to understand and generate content relevant to both domains. Eva-CLIP ViT [97] focuses on refining the extensive training process of its predecessor, CLIP. It aims to stabilize training and optimize performance, making the development of multimodal models more efficient and effective. For video content [98], a uniform sampling strategy can extract 5 frames from each video, applying similar preprocessing techniques as those used for images to ensure consistency in encoding and analysis across different media types.

Audio Modality: CFormer [99], HuBERT [100], BEATs [101], and Whisper [102] are key models for encoding audio, each with distinct mechanisms. CFormer integrates the CIF alignment method and a Transformer for audio feature extraction. HuBERT, inspired BERT [12], employs self-supervised learning to predict hidden speech units. BEATs focus on learning bidirectional encoder representations from audio via Transformers, showcasing advancements in audio processing.

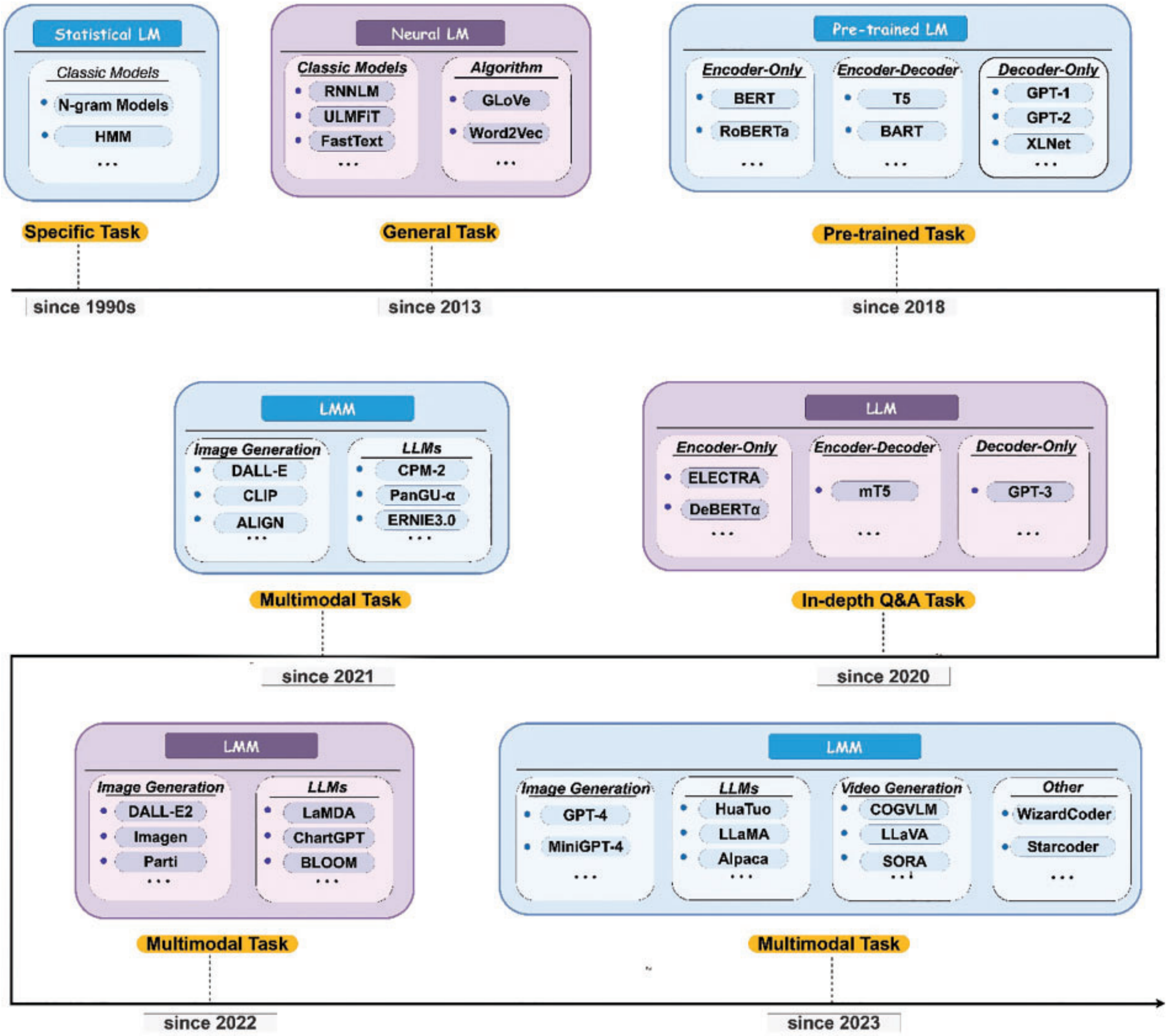

Figure 4: Images show evolving model tasks; multimodal models' adoption grows

2. *LLMs backbone*

LLMs backbone, as the central elements of LLMs, adopts key features such as zero-shot generalization, fewshot ICL (In-Context Learning), chain-of-thought (CoT), and adherence to instructions. The backbone of these LLMs manages to process and interpret representations across different modalities, facilitating semantic comprehension, logical reasoning, and input-based decision-making. In LMMs, frequently utilized LLMs encompass ChatGLM [55], FlanT5 [103], Qwen [83], Chinchilla [63], OPT [61], PaLM [28], LLaMA [30] and Vicuna [104], among others.

2.4.2 *Multimodal Generation*

The Modality Generator MGX plays a crucial role in generating outputs across different modalities. For this purpose, it often employs readily available Latent Diffusion Models (LDMs) [105], such as Stable Diffusion [106] for crafting images, Zeroscope [107] for creating videos, and AudioLDM-2 [108] for producing audio. The conditional inputs for this denoising process, which facilitates the creation of multimodal content, are provided by the features, as determined by the Output Projector. Hou et al. [109] developed a methodology that integrates image inputs and prompt engineering in LMMs to solve parson's problems, a kind of visual programming challenge.

2.5 Training Methods, Training Data and Testing

2.5.1 Training Methods

LLMs undergo a comprehensive development process encompassing pre-training, fine-tuning, and alignment to ensure their efficacy and ethical application across diverse tasks. Initially, pre-training equips LLMs with a broad understanding of language by learning from extensive text corpora, enabling them to capture complex linguistic patterns and knowledge. Subsequently, fine-tuning adjusts these pretrained models to specific tasks or domains, enhancing their performance on particular applications through targeted training on smaller, task-specific datasets. Finally, the alignment phase involves refining the models to adhere to human values and ethical standards, often employing techniques like Reinforcement Learning from Human Feedback (RLHF).

LMMs have significantly advanced the field of NLP by integrating text with other data modalities, such as images and videos. Presently, the development of LMMs follows three primary approaches: pretraining, instruction tuning, and prompting. In the forthcoming discussion, we will delve into these key strategies in greater detail.

1. *Pre-Training*

The pre-training process of LLMs is a pivotal step where models are exposed to vast amounts of textual data, enabling them to learn complex patterns, grammar, and contextual cues. Techniques such as Masked Language Modeling (MLM), prominently featured in BERT [12], involve obscuring parts of the text to challenge the model to infer the missing words using surrounding context. On the other hand, models like GPT leverage an autoregressive approach, predicting the next word in a sequence based on the words that precede it. This extensive pre-training phase allows LLMs to acquire a deep, nuanced understanding of language, facilitating their effectiveness across a broad spectrum of NLP tasks. The acquired knowledge enables these models to excel in applications ranging from text generation and summarization to question answering and translation, significantly advancing the field of AI and its capabilities in understanding and generating human language.

In the realm of LMMs, a significant trend is the integration of multiple modalities through end-to-end unified models. For example, MiniGPT-4 [76] leverages a pretrained and frozen ViT [98] alongside Q-Former and Vicuna LLMs [108], requiring only a linear projection layer for aligning vision and language modalities. Similarly, BLIP2 [79] introduces a dual-phase approach for vision-language modality alignment, starting with representation learning from a static visual encoder and progressing to vision-to-language generative learning facilitated by a static LLMs for zero-shot image-to-text tasks. Flamingo [15] further exemplifies this stream by utilizing gated cross attention mechanisms to merge inputs from a pre-trained visual encoder and an LLMs, effectively bridging the gap between visual and linguistic data.

2. *Fine-Tuning*

Instruction tuning of LLMs is a process designed to enhance models' ability to comprehend and execute textual instructions. This method involves training LLMs on datasets comprised of instructional prompts paired with corresponding outputs, thereby teaching the models to follow explicit directives. Such an approach significantly improves the model versatility, enabling it to perform a broad array of tasks as directed by user inputs. A prime example of this is GPT-3 [27], which has undergone instruction tuning to better understand and respond to natural language instructions. This enhancement allows GPT-3 to generate text that is not only relevant and coherent but also aligned with the specific instructions provided, showcasing its improved capacity for tasks ranging from content creation to answering complex queries. The success of instruction tuning in GPT-3 highlights its potential to make LLMs more interactive and adaptable, marking a significant advancement in the field of artificial intelligence and NLP.

Building on the concept of instruction tuning [107] for NLP tasks [110,111] researchers have expanded the scope to include fine-tuning pre-trained LLMs with multimodal instructions. This advancement enables the transformation of LLMs into multimodal chatbots [76,16,112] and task solvers [113–115] with notable examples being MiniGPT-4, BLIP2, and Flamingo for chatbots, and other models such as LaVIN and LLaMA Adapter focusing on task solving. A critical aspect of enhancing these LMMs involves gathering data that follows multimodal instructions for finetuning [116]. To overcome challenges associated with data collection, strategies such as benchmark adaptation [117–119] self-instruction [120–122], and hybrid composition [123,115] have been adopted. Furthermore, to bridge the modality gap, a learnable interface connects different modalities from frozen pre-trained models, aiming for parameter-efficient tuning. For instance, LaVIN [123] and LLaMA Adapter [124] have introduced transformer-based and modality-mixing adapter modules, respectively, for efficient training. In contrast, expert models like VideoChatText [17] leverage specialized models such as Whisper [102] for speech recognition, converting multimodal inputs directly into language, thereby facilitating comprehension by subsequent LLMs.

3. *Alignment*

The concept of alignment pertains to the process of aligning the model outputs with human values, intentions, and ethical standards. This involves training methodologies and evaluation strategies designed to ensure that LLMs behave in ways that are beneficial and non-harmful. Wei et al. [125] introduce techniques for aligning LLMs through iterative processes involving human feedback, where models are fine-tuned based on evaluations of their outputs against desired ethical and moral criteria. Furthermore, Ouyang et al. [110] explore alignment through RLHF, a method where models are adjusted based on direct human input on the appropriateness and alignment of generated content. These processes aim to mitigate risks associated with LLMs generating biased, misleading, or harmful content, ensuring their utility and safety in real-world applications.

4. *Prompting*

Prompting techniques, in contrast to fine-tuning, offer a way to guide Mega LMMs using context or instructions without changing their parameters, reducing the need for vast multimodal datasets. This method is particularly useful for multimodal CoT tasks, allowing models to generate reasoning and answers from multimodal inputs. Examples include CoT-PT [126], which uses prompt tuning and visual biases for implicit reasoning, and Multimodal-CoT, employing a two-step process combining rationale generation and answer deduction. This approach also facilitates breaking down complex tasks into simpler sub-tasks through multimodal prompts [24,127], demonstrating the effectiveness and adaptability of prompting in multimodal learning.

2.5.2 *Data Sources and Evaluation*

1. *Datasets for LLMs*

LLMs derive their prowess from meticulously curated datasets, which are essential for their pre-training and finetuning. The creation of such datasets is a demanding task, demanding both breadth and depth of high-quality data. Researchers have identified an array of key data sources integral for LLMs training, as summarized in Table 2. This selection includes literature, dialogues, code from GitHub, comprehensive common crawl data, domain-specific datasets from NLP tasks, academic content from Stanford University, and discussions from Reddit. Additionally, synthetic data and the expansive knowledge from Wikipedia are harnessed, providing LLMs with a rich, varied linguistic and conceptual landscape to learn from, thereby enhancing their applicability across a myriad of tasks. The specific attributes of these datasets are systematically detailed in Table 3.

Table 3: Sources of pre-training and fine-tuning datasets for LLMs

Data sources	Dataset	Size	Type
Books	BookCorpus [128]	5 GB	Pre-training
	Gutenberg [129]	–	Pre-training
Chat	HH-RLHF [130]	160 K	Fine-tuning
	Dolly [131]	15 K	Fine-tuning
	HC3 [132]	87 K	Fine-tuning
	OpenAssistant [133]	161 K	Fine-tuning
	ShareGPT [134]	90 K	Fine-tuning
Codes	BigQuery [135]	–	Pre-training
Common crawl	C4	800 GB	Pre-training
	CC-NEWS [91]	78 GB	Pre-training
	CC-Stories-R [136]	31 GB	Pre-training
	mC4 [68]	38.49 TB	Pre-training
	REALNEWs [137]	120 GB	Pre-training
GitHub	BigPython [70]	5.5 TB	Pre-training
NLP task	FLAN	4.4 M	Fine-tuning
	MVPCorpus [138]	41 M	Fine-tuning
	Nat. Inst. [139]	193 K	Fine-tuning
	OIG [140]	43 M	Fine-tuning
	P3 [141]	12.1 M	Fine-tuning
	Super Nat. Inst. [142]	5 M	Fine-tuning
	xP3 [143]	81 M	Fine-tuning
Reddit links	OpenWebText [144]	38 GB	Pre-training
	Pushift.io [145]	2 TB	Pre-training
Stanford University	CoQA [64]	–	Fine-tuning
Synthetic	Alpaca [146]	52 K	Fine-tuning
	Baize [147]	158 K	Fine-tuning
	BELLE [148]	1.5 M	Fine-tuning
	Guanaco [149]	535 K	Fine-tuning
	Self-Instruct [120]	82 K	Fine-tuning

(Continued)

Table 3 (continued)

Data sources	Dataset	Size	Type
Webpages	RefinedWeb [150]	2.8 TB	Fine-tuning
Wikipedia	Wikipedia [151]	21 GB	Pre-training
Other	Infiniset [53]	–	Pre-training
	MassiveText [62]	10.5 TB	Pre-training
	RedPajama [152]	2.7 TB	Pre-training
	ROOTS [153]	1.6 TB	Pre-training
	the Pile [154]	800 GB	Pre-training
	The Stack [155]	6 TB	Pre-training
	LIMA [156]	1 K	Fine-tuning
	OPT-IML [111]	18.1 M	Fine-tuning
	SNLI [157]	570 K sentence-pairs	Fine-tuning
	SQuADv2 [158]	44 MB	Fine-tuning

Pre-Training Datasets: Pre-training datasets serve as the foundation for the development of LLMs, and their diversity is crucial for the comprehensive understanding these models achieve. For instance, the extensive common crawl dataset captures a wide snapshot of the web, enabling models to learn from a myriad of topics and writing styles. Similarly, literary works included in datasets provide nuanced language and complex narrative structures that aid in understanding more sophisticated language use. GitHub repositories contribute technical and programming language data, essential for specialized tasks as discussed by Zhao et al. [105]. Meanwhile, the Reddit dataset, with its conversational and often informal text, offers insights into colloquial language. These datasets, among others, are instrumental in pre-training LLMs, equipping them with the breadth of knowledge necessary to understand and generate humanlike text.

Instruction-Tuning Datasets: Following the pre-training phase, instruction tuning, also known as supervised finetuning, plays a crucial role in amplifying or eliciting particular competencies in LLMs. We delve into a selection of prominent datasets employed for instruction tuning, which we have organized into three principal categories according to how the instruction instances are formatted. These categories encompass datasets oriented towards NLP tasks, datasets derived from everyday conversational interactions, and artificially generated, or synthetic, datasets.

Alignment Datasets: Beyond instruction tuning, crafting datasets that ensure LLMs align with human ethical standards, such as helpfulness, truthfulness, and nonmaleficence, is crucial. This section presents a suite of key datasets employed for alignment tuning. The statistical table for the specific datasets used for pre-training data and instruction tuning data of LLMs can be found in Table 4.

Table 4: Datasets for alignment in LLMs

Dataset	Size	Type
Anthropic-HH-RLHF [130]	142 K	Alignment
Anthropic-HH-RLHF-2 [159]	39 K	Alignment

(Continued)

Table 4 (continued)

Dataset	Size	Type
CValues [160]	145 K	Alignment
PKU-SafeRLHF [161]	330 K	Alignment
Sandbox alignment data [162]	169 K	Alignment
SHP [163]	385 K	Alignment
Stack exchange preferences [164]	10 M	Alignment
Summarize from feedback [165]	193 K	Alignment
WebGPT comparisons [166]	19 K	Alignment
CB	–	Evaluation

2. Datasets for LMMs

LMMs leverage vast and varied datasets for pre-training, encompassing image, text, and sometimes audio-visual content to understand and generate across modalities. Pretraining on datasets and BooksCorpus [128] for text allows LMMs to acquire foundational knowledge. Instruction tuning datasets then tailor these models for specific tasks; for instance, the visual question answering dataset guides models on how to respond accurately to queries about visual content. Such comprehensive training enables MM-LMs to perform complex tasks like image captioning and visual reasoning, bridging the gap between human and machine perception. Evaluating LMMs encompasses measuring their proficiency in tasks combining text and visual inputs. This involves specialized benchmarks, aiming to quantify the models understanding and generative capabilities across modalities. These evaluations are crucial for gauging how well MMLMs can mimic human-like understanding in diverse scenarios. For a detailed overview of the datasets employed in these evaluations, refer to Table 5 and Table 6.

Table 5: Sources of pre-training and fine-tuning datasets for LMMs

Data sources	Dataset	Size	Type
GitHub	COYO-700 M [167]	747 M	Pre-training
Google	ALIGN [77]	1.8 B	Pre-training
	CC12 M [168]	12.4 M	Pre-training
	CC3M [169]	3.3 M	Pre-training
	JFT-300M [170]	300 M	Pre-training
	JFT-3B [171]	3 B	Pre-training
Microsoft	MS-COCO [172]	620 K	Pre-training
MS-COCO	COCO Caption [173]	1 M	Pre-training
Stanford University	GQA [174]	22 M	Fine-tuning
Other	OCR-VQA [175]	1 M	Fine-tuning
	Pathvqa [176]	–	Fine-tuning
	QuAC [177]	100 K	Fine-tuning
	RefCOCO [178]	142 K	Fine-tuning

(Continued)

Table 5 (continued)

Data sources	Dataset	Size	Type
	RefCOCO+ [179]	142 K	Fine-tuning
	Slake [180]	14KQA	Fine-tuning
	ST-VQA [181]	32 K	Fine-tuning
	TextVQA [182]	45.3 K	Fine-tuning
	VGQA [183]	1.7 M	Fine-tuning
	Visual-7W [184]	328 K	Fine-tuning
	VQA-RAD [185]	3KQA pairs	Fine-tuning
	VQAv2 [186]	1.4 M	Fine-tuning
	A-OKVQA [187]	24.9 K	Fine-tuning
	DataComp [188]	1.4 B	Fine-tuning
	DocVQA [189]	50 K	Fine-tuning
	DVQA [190]	3.5 M	Fine-tuning
	RedCaps [191]	–	Pre-training
	SBU [192]	1 M	Pre-training
	Visual Genome [183]	4.5 M	Pre-training
	WIT [193]	37 M + image-text	Pre-training
	YFCC [194]	–	Pre-training
	Ai challenger [195]	1.5 M	Pre-training
	Flickr30K [196]	158 K	Pre-training
	Flickr30k entities [197]	–	Pre-training
	IG-3.6 B [198]	3.6 B	Pre-training
	ImageNet [199]	–	Pre-training
	ImageNet-1K [200]	1.2 M	Pre-training
	ImageNet-21K [201]	14 M	Pre-training
	LAION-2B [202]	–	Pre-training
	LAION-400M [203]	400 M	Pre-training
	LAION-5B [202]	5.9 B	Pre-training
	LAION-COCO [204]	600 M	Pre-training
	LAION-en [202]	2.3 B	Pre-training
	LAION-zh [202]	142 M	Pre-training

Table 6: Datasets for evaluation, specialization, and other purposes in LMMs

Dataset	Size	Type
NoCaps [205]	–	Evaluation
SEED [206]	–	Evaluation
VSDial-CN	1.2 M	Evaluation
VTP [15]	27 M	Evaluation
WaveCaps [19]	403 K	Evaluation
WebVid [207]	10 M	Evaluation

(Continued)

Table 6 (continued)

Dataset	Size	Type
ScienceQA [208]	–	Evaluation
OBELISC [209]	468 B	Evaluation
MSRVTT [210]	200 K	Evaluation
Text Captions [211]	145 K	Specialized
WebLI [212]	12 B	Specialized
Wukong [213]	101 M	Specialized
CC595k [45]	595 K	Specialized
Episodic WebLI [214]	400 M	Specialized
MMC4 (Interleaved) [215]	101.2 M (Instances)	Other
Aishell-2 [216]	1 M/128 K	Other
I2E [217]	1.1 B	Other
LTIP [15]	312 M	Other
M3W (Interleaved) [15]	43.3 M (Instances)	Other

Addressing diversity and mitigating potential biases in large datasets are essential tasks in the development of AI models. The following strategies are implemented to ensure that the data used does not perpetuate or amplify bias:

Fairness Metrics: Various fairness metrics are used to evaluate AI models to ensure they do not favor one group over another [218]. These metrics like ImageNet [219] for images help in understanding and quantifying any disparities in model performance across different groups defined by attributes like age, gender, ethnicity, etc.

Bias Detection and Mitigation: Specialized tools and methodologies are used to detect and quantify biases in datasets. Once identified, strategies such as re-sampling the data, weighting, or modifying the data processing techniques are employed to mitigate these biases.

Regular Audits: Periodic audits of the AI models and their training data help in identifying and addressing any emergent biases or issues in performance. These audits are crucial for maintaining the integrity and fairness of the model over time.

Each row within the Table 7 meticulously outlines the model's name, parameter count, number of layers, dataset descriptions, and their respective training strategies, including autoencoding methods, autoregressive methods and sequence-to-sequence (Seq2Seq) encoding-decoding methods. A comparative overview of several large models, including parameter sizes, layers, datasets, and training regimes (the "-" indicates that for multimodal models, due to their unique architectures and methods of integrating various types of data, certain details such as the number of layers or training strategies are not readily classifiable or applicable, hence these fields are left blank). The blue bottom represents the LLMs, and the red bottom represents the LMMs. This comprehensive summary facilitates a deeper understanding of the diversity and scale of contemporary language models, as well as the complexities associated with their data processing and learning mechanisms.

Table 7: Comparative overview of LLM and LMM parameters, layers, datasets, training regimes

Models	Params	*n* Layers	Dataset	Training regimes
BERT [12]	340 M	24	**BooksCorpus:** a dataset consisting of a large collection of book texts. **English Wikipedia:** the English version of Wikipedia, containing vast encyclopedic knowledge. **Common Crawl:** covering a wide range of topics and languages.	Encoder-only (Autoencoding)
T5 [13]	11 B	24	**Colossal Clean Crawled Corpus (C4):** an unlabeled pure English dataset, abbreviated as C4, is approximately 750 GB in size.	Encoder-decode (Seq2Seq)
Bart [88]	406 M	12	–	
ChatGLM3-6B [105]	6 B	96	The dataset refers to the dataset of BERT.	Autoregressive blank filling
ERNIE3.0 [46]	280 M	48	11 large-scale, multi-variety, high-quality Chinese text corpora in different categories with a capacity of up to 4 TB.	Encoder-decode (Seq2Seq)
PanGu-α [47]	200 B	64	The dataset of Common Crawl after data cleansing.	Decode-only (Autoregressive)
BLOOM [51]	176 B	70	**ROOTS:** a corpus consisting of 498 Hugging Face datasets. A total of 1.61 TB of text, including 46 natural languages and 13 programming languages.	Decode-only (Autoregressive)
CPM-2 [48]	11 B	24	Encyclopedias, novels, Q&As, scientific literature, e-books, news and reviews, etc. From the 50 TB of raw data, 2.3 TB of Chinese data and 300 GB of English data were cleaned.	Encoder-decode (Seq2Seq)

(Continued)

Table 7 (continued)

Models	Params	*n* Layers	Dataset	Training regimes
OPT [61]	175 B	96	A subset of the **RoBERTa corpus** and Stories, and a newer version of CCNews. A subset of the **Pile corpus** CommonCrawl, DM Mathematics, Project Gutenberg, HackerNews, OpenSubtitles, OpenWebText2, USPTO, and Wikipedia.	Decode-only (Autoregressive)
Galactica [29]	120 B	96	Papers, Code, Reference Material, Knowledge Bases, Filtered CommonCrawl, Prompts, Other.	Decode-only (Autoregressive)
LaMDA [53]	137 B	64	2.97 B documents, 1.12 B dialogs, and 13.39 B dialog utterances, for a total of 1.56 T words.	Decode-only (Autoregressive)
PaLM [28]	540 B	118	Social media conversations, filtered webpages, books, GitHub, Wikipedia, News.	Decode-only (Autoregressive)
LLaMA2 [30]	70 B	80	2 trillion tokens of data from publicly available sources.	Decode-only (Autoregressive)
Qwen [83]	3 T	–	Publicly available documents on the web, encyclopedias, books, code repositories, among others.	–
Gemini [220]	1560 B	–	Publicly available documents on the web, code, including images, audio, and video data.	–
DALL-E [72]	12 B	–	A dataset of 250 million image-to-text pairs.	–
ALIGN [77]	800 M	–	A dataset of 18 billion image-to-text pairs.	–
CLIP [78]	400 M	–	A dataset of 400 million image-to-text pairs.	–

In summary, dealing with diversity and potential bias in large datasets is a critical task in AI model development. To ensure that the data used does not convey or amplify these biases, the RD team uses a variety of strategies. First, they conduct a data review to identify and correct biased data. Second, enhance the diversity of the dataset by sourcing data from a variety of origins to guarantee equitable representation across different demographics. In addition, developers will use algorithm review and testing to ensure that the model's decision-making process is fair and unbiased. These methods work together to help build more unbiased and reliable AI systems.

LLMs typically utilize deep neural network architectures, such as the Transformer architecture, which comprises a multi-layer self-attention mechanism and a feed-forward neural network. Each layer processes input data through encoding and decoding steps. The number of layers directly influences the model's complexity and performance. For instance, GPT-3 [27], with its 175 billion parameters, features a 96-layer Transformer architecture. In contrast, multimodal models handle not only text data but also integrate diverse types of data like images, audio, and video. These models necessitate specialized layer structures to process and fuse multiple data modalities. For example, the CLIP [78] model combines images and text to learn cross-modal representations through parallel visual and text Transformers.

The parameter scales of both LLMs and LMMs are immense. Increasing parameter size generally enhances performance but also escalates the demand for computational resources and training complexity. Training LLMs typically depends on vast text datasets collected from the internet, including books, articles, and website content. For instance, BERT [12], which possesses 340 million parameters, utilize datasets like the BooksCorpus and English Wikipedia.

Similarly, LMMs often possess large parameter sizes due to the simultaneous processing and fusion of multiple data modalities. To be specific, LLaMA2 [30] boast significantly higher parameters, approximately 70 billion, harnessing an extensive amount of data—2 trillion tokens—from publicly available sources, and are trained using a decode-only approach.

LMMs require datasets that encompass various data types such as images and text. For instance, Gemini [220], which holds 1560 billion parameters and leverages public documents across diverse formats, including images and audio.

3. Evaluation

In the evaluation strategy of LLMs and LMMs, three pivotal methodologies have emerged to assess their performance and capabilities comprehensively: the benchmark-based approach, the human-based approach, and the model-based approach. Each of these methods offers distinct advantages and inherent limitations, often necessitating their combined application for a thorough appraisal of an LLMs' and LMMs' proficiency.

Benchmark-based approach: In evaluating LLMs, segmenting benchmarks into knowledge-centric and reasoning-centric categories. Knowledge benchmarks like MMLU [221] and CEval [222] are designed to measure the models' grasp of factual information, while reasoning benchmarks such as GSM8K [223], BBH [224], and MATH [225] evaluate their ability to engage in complex problem-solving. The evaluation process entails generating responses from LLMs to structured prompts and then employing a set of rules to predict answers from these responses. The accuracy of the models is quantified by comparing these predictions to the correct answers. In the realm of LMMs, there has been a concerted effort to develop benchmarks tailored to their unique capabilities. Notably, Fu et al. [226] developed the MME benchmark, a suite that encompasses 14 distinct perceptual and cognitive tasks, with each instruction-answer pair meticulously crafted to prevent data leakage.

Additionally, the LAMM-Benchmark [227] was introduced for the quantitative assessment of LLMs across a spectrum of 2D and 3D visual tasks. Video-ChatGPT [228] has also contributed to this space by presenting a framework for evaluating video-based conversational models, including assessments of video-based generative performance and zero-shot question-answering capabilities. These advancements in benchmarking are crucial for the thorough evaluation and continuous refinement of LLMs and LMMs.

Human-based approach: The human-based approach to evaluating LLMs is pivotal for assessing real-world applicability, including alignment with human values and tool manipulation. This method uses open-ended questions, with human evaluators judging the quality of LLM responses. Employing techniques like pairwise comparison and single-answer grading, evaluations can range from direct answer scoring in HELM [229] for tasks like summarization to comparative feedback in Chatbot Arena's [230] crowdsourced conversations. This nuanced assessment is essential for tasks requiring humanlike judgment and creativity, providing a comprehensive view of LLMs capabilities. Evaluating LMMs as chatbots involves open-ended interactions, challenging traditional scoring methods. Assessment strategies include manual scoring by humans on specific performance dimensions. While manual scoring is insightful, it is labor-intensive.

Model-based approach: The model-based approach offers a promising solution to labor-intensive problems. GPT scoring, leveraging models such as GPT-4, is utilized to evaluate responses based on criteria like helpfulness and accuracy. However, this method encounters limitations due to the non-public availability of multimodal interfaces, potentially impacting the accuracy of performance benchmarks. Furthermore, case studies are conducted to provide a more detailed analysis, particularly beneficial for complex tasks requiring sophisticated human-like decision-making. In evaluating LLMs, additional models or algorithms are employed to assess their performance, revealing both intrinsic capabilities and shortcomings. To mitigate the high cost associated with human evaluation, surrogate LLMs like ChatGPT and GPT-4 are utilized. Platforms such as AlpacaEval [231] and MT-bench [232] employ these surrogate LLMs for comparative analysis. While these closed-source LLMs exhibit high concordance with human assessments, concerns persist regarding access and data security. Recent endeavors have concentrated on fine-tuning open-source LLMs, such as Vicuna [104], to function as evaluators, thereby narrowing the performance gap with proprietary models.

2.6 Emergent Abilities of LLMs

Wei et al. [233] studied the emergence abilities of large-scale language models, a phenomenon does not present in smaller models. As these models increase in size, they develop new, unpredictable capabilities that surpass the performance of smaller models, akin to phase transitions in physics [234]. While emergent abilities can be task-specific [62], the emphasis here is on versatile abilities that enhance performance across diverse tasks. This part introduces three principal emergent abilities identified in LLMs, alongside models that demonstrate such capabilities [235].

2.6.1 In-Context Learning (ICL)

ICL was notably defined in the context of GPT-3 [27], illustrating that when provided with natural language instructions and/or task demonstrations, the model can generate accurate outputs for test instances by completing input text sequences, without necessitating further training or adjustments. This capability, particularly pronounced in the GPT-3 model with 175 billion parameters, was not as evident in earlier iterations such as GPT-1 and GPT-2. However, the effectiveness of ICL varies with the nature of the task at hand. For instance, GPT-3, 13 billion parameter variant demonstrates proficiency in arithmetic tasks, like 3-digit addition and subtraction, whereas the more extensive 175 billion parameter model struggles with tasks like Persian question answering.

2.6.2 Instruction Following

Instruction tuning, which involves fine-tuning LLMs with a diverse set of tasks described through natural language, has proven effective in enhancing the model's ability to tackle novel tasks also presented in instructional form. This technique allows LLMs to understand and execute instructions for new tasks without relying on explicit examples [116,144,223], thereby broadening their generalization capabilities. Research indicates that the LaMDA-PT model [53], after undergoing instruction tuning, began to markedly surpass its untuned counterpart in performing unseen tasks at a threshold of 68 billion parameters, a benchmark not met by models sized 8 billion parameters or less. Further studies have identified that for PaLM [28] to excel across a range of tasks as measured by four evaluation benchmarks (namely MMLU, BBH, TyDiQA, and MGSM), a minimum model size of 62 billion parameters is necessary, although smaller models may still be adequate for more specific tasks, such as those in MMLU [107].

2.6.3 Step-by-Step Reasoning

Zhou et al. [236] proposed a novel NLP technique, "From Simple to Complex Prompting", which breaks down complex problems into simpler sub-problems, starting with the easiest and progressively tackling to more challenging ones. This method is particularly effective for complex issues, outperforming traditional methods. The CoT prompting strategy allows LLMs to address tasks through a prompting mechanism that incorporates intermediate reasoning steps towards the final solution [31,237].

3 Open Issues

This section embarks on an exploration of the yet unresolved complexities inherent in LLMs and LMMs. These advanced computational systems, while demonstrating unprecedented abilities in processing and generating human-like text and multimedia content, still grapple with significant challenges. We delve into the intricate nuances of these models, examining the limitations that hinder their full potential. Key areas of focus include the ongoing struggle with understanding and replicating nuanced human context, the management of inherent biases in training data, and the challenges in achieving true semantic understanding.

3.1 Contextual Understanding

Contextual understanding is a hallmark capability of both LLMs and LMMs, revolutionizing the way information is processed and interpreted across various domains. LLMs excel in comprehending and generating text within specific contexts, discerning nuances and subtleties to produce coherent and contextually appropriate responses. Similarly, LMMs extend this prowess by incorporating diverse modalities such as images, audio, and text, allowing for a richer understanding of complex scenarios. Whether it's analyzing textual documents or interpreting visual cues alongside linguistic context, both LLMs and LMMs demonstrate a remarkable ability to grasp and interpret the intricate interplay of contextual factors, thereby advancing research, problem-solving, and decision-making across diverse fields.

3.1.1 Contextual Limitations

In the realm of advanced computational linguistics, both LLMs and LMMs for Matching encounter a critical challenge known as contextual limitations. Chen et al. [238] introduced Position Interpolation (PI) as a solution to the limitations posed by insufficient context window sizes in models. The essence of this approach lies in avoiding extrapolation. Instead, it focuses on reducing position indices by aligning the maximum position index with the context window upper limit, as set during pre-training. This alignment of the position index range and relative distances before and after expansion mitigates the effects of expanding the context window on attention score calculation. Consequently, it enhances the model adaptability while preserving the quality associated with the original context window size. Exploration of methodologies to enrich contextual reasoning capabilities is imperative, facilitating models to deduce implicit information and formulate nuanced predictions grounded in comprehensive contexts. This endeavor may entail delving into sophisticated techniques, such as integrating external knowledge reservoirs or harnessing multi- hop reasoning mechanisms.

3.1.2 Ambiguity and Vagueness

This section discusses the often-encountered issues of ambiguity and vagueness in contextual understanding, analyzing relevant research and proposing strategies to address them. As an illustration, Chuang et al. [239] proposed a new decoding method called Decoding by Contrasting Layers (DoLa). This method seeks to enhance the extraction of factual knowledge embedded within LLMs without relying on external information retrieval or additional fine-tuning. Capitalizing on the observation that factual knowledge in LLMs is often confined to specific transformer layers, DoLa derives the next label distribution by comparing the logarithmic differences obtained through projecting the front and back layers into the vocabulary space. Concretely, it involves subtracting the logarithmic probability of the output from the mature layer from the output of the immature layer. This resultant distribution is then employed as the prediction for the next word, with the overarching goal of minimizing ambiguity and addressing other related challenges. As another important aspect, unlike enhancing the extraction of internal knowledge in LLMs, extracting valuable evidence from the external world allows for answering questions based on the gathered evidence [240–243]. Specially, LLM-Augmenter [244] are proposed to enhance the performance of LLMs. In contrast to a standalone LLM, it introduces a set of plug-and-play modules, enabling the LLMs to leverage external knowledge for generating more accurate and information-rich responses. The system continuously optimizes the LLMs prompts based on feedback generated by utility functions, enhancing the quality of the models' responses. In a range of settings, including task-focused conversations and broad-spectrum query response systems, the LLM-Augmenter efficiently minimizes the generation of spurious outputs by the LLMs, all the while preserving the response coherence and richness of information. Furthermore, based on a multimodal LLMs framework, Qi et al. [245] introduced a systematic approach to probing multimodal LLMs using diverse prompts to understand how prompt content influences model comprehension. It aims to explore the model's capability through different prompt inputs and assess contextual understanding abilities with a series of probing experiments. Existing research has explored the inconsistency between vision and language. For instance, Khattak et al. [246] proposed a novel method addressing the inconsistency between visual and language representations in pre-trained visual language models like CLIP. It enhances collaborative learning by integrating multimodal prompts into both vision and language branches, thereby aligning their outputs. The method employs cross-entropy loss for training and has been evaluated across 11 recognition datasets, consistently outperforming existing methods.

Future work on ambiguity and vagueness entails several pivotal avenues for advancement in natural language understanding [247]. Primarily, researchers aim to develop robust algorithms capable of effectively disambiguating ambiguous terms and resolving vague expressions within textual contexts. This involves exploring novel techniques such as context-aware word sense disambiguation and probabilistic modeling of vague language.

Additionally, further investigation is warranted to enhance the capacity of models to handle inherent ambiguities and vagueness in human language. This could involve the development of advanced machine learning approaches that integrate contextual information and domain knowledge to make more informed interpretations of ambiguous or vague statements.

3.1.3 Catastrophic Forgetting

Catastrophic forgetting in LLMs and LMMs is a critical challenge. It arises when these models, after being trained or fine-tuned on new data or tasks, tend to forget the knowledge they previously acquired. This issue occurs because the neural network weights, which are adjusted to improve performance on new tasks, might overwrite or weaken the weights essential for earlier tasks. This problem is especially acute in LLMs and LMMs due to their intricate structure and the vast variety of language data they process. It significantly hinders their ability to consistently perform across different tasks, particularly in dynamic settings that demand continuous learning and adaptation.

In particular, Mitra et al. [248] proposed a novel approach for improving the performance of LMMs in vision-language tasks, named Compositional Chain-of-Thought (CCoT) to address the issues of the forgetting of pre-training objectives. CCoT operates in two primary steps. First, a scene graph is generated using an LMM, which involves creating a structured representation of the visual scene. The second step involves using the generated scene graph as part of a prompt in conjunction with the original image and task prompt. Additionally, by incorporating scene graphs, CCoT allows for a more organized and comprehensive processing of visual information. In order to overcome the same problem in the field of image reasoning, BenchLMM [249] is first utilized to assess the performance of LMMs across various visual styles, addressing the issue of performance degradation under non-standard visual effects. After image processing, PixelLM [250] excels in creating detailed object masks, addressing a key shortfall in multimodal systems. Its core includes a novel lightweight pixel decoder and segmentation codebooks, streamlining the transformation of visual features into precise masks. This innovation enhances task efficiency and applicability in areas like image editing and autonomous driving. Additionally, the mechanism for target refinement loss in the model improves discrimination of overlapping objectives, thus refining mask quality.

Besides the above methods, Liu et al. [251] introduced the DEJAVU system to improve the efficiency of LLMs during inference, addressing the high computational cost issue without sacrificing contextual learning abilities. Unlike existing methods that require costly retraining or reduce LLMs contextual capabilities, DEJAVU dynamically forecasts contextual sparsity based on input data for each layer, combined with asynchronous processing hardware implementation. This approach significantly reduces inference latency, outperforming prevalent systems like FasterTransformer and hugging face implementations.

Future research on catastrophic forgetting encompasses several critical areas aimed at mitigating this phenomenon and enhancing the robustness of neural networks in continual learning scenarios. Researchers are exploring methods to design neural architectures that are more resistant to catastrophic forgetting, such as incorporating mechanisms to selectively retain important information from previous tasks while learning new ones.

To be specific, there is a need to develop more effective rehearsal-based learning techniques, where models actively revisit and train on past data to prevent forgetting. This may involve investigating strategies for prioritizing and sampling past experiences in a way that maximally benefits learning on new tasks.

3.2 Hallucination Correction and Cognitive Ability Evaluation

LLMs and LMMs, while being marvels of modern computational linguistics, are not exempt from an intriguing phenomenon known as “hallucinations”—where outputs generated by the model are either factually incorrect or nonsensical. This phenomenon primarily arises from several core issues. First and foremost, the quality and scope of the training data play a pivotal role. Secondly, the model limitations in understanding context led to hallucinations. Additionally, the challenge of reasoning and common sense is also apparent. LLMs, adept at pattern recognition and language generation, sometimes falter in tasks requiring logical reasoning or common-sense knowledge, resulting in responses that seem plausible but are fundamentally flawed. Another contributing factor is the inherent limitations of the model architectures and algorithms.

3.2.1 Corrective Methods and Evaluation

In the evolving landscape of artificial intelligence, the phenomena of “hallucination” in LLMs and LMMs present a unique set of challenges and opportunities. This section of the paper delves into the intricate world of hallucination correction within these advanced AI systems. It explores the mechanisms through which these models occasionally generate misleading or factually incorrect information, often in response to complex or ambiguous prompts. The focus then shifts to the evaluation of cognitive abilities in AI, scrutinizing how these systems understand, process, and respond to diverse information. By dissecting the underpinnings of hallucination and assessing the cognitive competencies of these models, this paper aims to shed light on the path forward in refining AI for more accurate, reliable, and contextually aware responses.

In terms of zero resource illusion recognition, SelfCheckGPT [252] is proposed to achieve zero-resource black-box hallucination identification within generative LLMs. The fundamental principle asserts that a language model, once it comprehends a specific concept, is expected to produce responses through random sampling. These responses should not only resemble each other but also uphold consistent truths.

Conversely, for hallucinated content, randomly selected replies are prone to divergence and contradictions. The research findings indicate that SelfCheckGPT effectively detects both non-factual and factual sentences and ranks the authenticity of the content. Compared to gray-box methods, this approach demonstrates superior performance in sentence-level hallucination detection and paragraph-level authenticity assessment.

Furthermore, Friel et al. [253] proposed the innovative Chain-Poll methodology and the RealHall benchmark suite as powerful tools for evaluating and solving the hallucinogen difficulty in LLMs outcomes, making a comprehensive and impactful contribution to the field of hallucinogen detection in LLM-generated texts. To be specific, the RealHall benchmark suite has been designed to address the limitations of previous hallucination detection efforts. ChainPoll is designed to detect both open and closed domain hallucinations, thus demonstrating its versatility. Performance tests conducted in this thesis show that ChainPoll outperforms a range of published alternatives, including SelfCheckGPT [252], GPTScore [254], G-Eval [255], and TRUE [256]. ChainPoll proves to be not only more accurate, but also faster, more cost-effective, and equally good at detecting both open and closed domain illusions.

In the realm of LLMs and LMMs, future research on corrective methods and evaluation encompasses several pivotal avenues for enhancing model performance and reliability. This includes exploring advanced techniques for model calibration and fine-tuning to rectify errors and biases in model outputs, as well as developing comprehensive evaluation metrics and benchmarks to accurately assess model performance in real-world scenarios. Additionally, there is a growing emphasis on incorporating feedback and corrective signals into the training process to facilitate continual improvement and adaptation of models over time, which may involve exploring active learning techniques and integrating human supervision into the training process. Overall, this future work aims to propel the field forward by refining model accuracy, improving evaluation methodologies, and enabling models to adapt effectively in dynamic environments.

3.2.2 Multimodal Methods and Research

In terms of visual and language tasks, there have also been studies introduced a novel video-audio zero-shot learning approach [257], leveraging multi-modal data alignment and a multi-channel attention mechanism for knowledge transfer. Incorporating datasets like VG Sound, UCF, and Activity Net, it tests its method against realistic scenarios and demonstrates its effectiveness. Recent studies, such as UMT [258], VL-ADAPTER [259], RLHF [260], have investigated the signifiance of multimodal learning in handling complex tasks. By comparing the performance and efficiency of models, they reached analogous conclusions: multimodal learning, through the amalgamation of visual, auditory, and textual inputs, can yield a more enriched and holistic representation of data. This integration significantly addressing the issue of multimodal misalignment and hallucinations in LMMs. In addition, it has been demonstrated that utilizing unique approaches, such as matrix-based feature extraction and adapter-based techniques, can effectively enhance the performance of models. This underscores the importance and potential of innovative methods in the field of multimodal learning. Therefore, as suggested in [261,262], it is essential to pay attention to the scale of the model and the quality of the input data.

3.2.3 Actual Impact

To be specific, computer illusions often trouble practical work. In the medical field, hallucinations produced by large models, can lead to issues like misdiagnosis, incorrect treatments, and privacy breaches. Umapathi et al. [263] proposed a new benchmark and dataset, Med HALT, designed to assess and reduce hallucinations in LLMs in the medical domain. The authors evaluated the performance of several leading LLMs on the Med-HALT dataset, including GPT-3.5, Davinci, Falcon 40B, and Llama-2 70B. revealing differences in their performance, they found that, while all of the models performed well on factual questions, they did not perform well on more complex reasoning and IR tasks. the Med-HALT dataset and benchmarks provide a valuable resource for assessing and improving the reliability and safety of LLMs in healthcare. The authors hope that their work will encourage further research and collaboration in the field and facilitate the pursuit of reproducible results.

The future prospects regarding the actual impact of LLMs and LMMs are promising and multifaceted. Researchers anticipate further advancements in these models leading to transformative effects across various domains and industries. These models are poised to revolutionize natural language understanding, image recognition, and audio processing, among other tasks, by offering increasingly accurate and versatile solutions.

However, alongside these opportunities, it is essential to address challenges related to bias, fairness, privacy, and ethical considerations in the deployment of these models. Future research and development efforts will need to focus on mitigating these risks and ensuring that the benefits of large language and multimodal models are equitably distributed and ethically sound.

3.3 Accuracy Reasoning

In accuracy reasoning, we will analyze the two parts of internal reasoning and external reasoning, mainly introducing complex internal reasoning mechanisms, thinking chains, and the use of external tools.

3.3.1 Internal Reasoning

In the intricate domain of artificial intelligence, LLMs and LMMs stand as beacons of innovation. These models, characterized by their extensive data assimilation and processing, are redefining the paradigms of machine cognition. Central to their groundbreaking capabilities is a complex internal reasoning mechanism. This exploration aims to unravel the enigmatic cognitive processes underpinning these advanced systems. We venture into their elaborate architectures, discerning how they transcend traditional computational roles to emerge as harbingers of a new epoch in digital intelligence. Li et al. [264] proposed a technique crafted to improve the "truthfulness" of LLMs, named Inference-Time Intervention (ITI). To be specific, ITI functions by altering the activations of the model during the inference process, following specific pathways across a select few attention heads. After this process, the complete answer is generated. ITI is a minimally invasive control method, that can leverage the potential knowledge of LLMs. Unlike existing methods, it does not require a large number of annotations and computational resources. But ITI cannot guarantee that LLMs always provides true answers, nor can it cover all the meanings of authenticity. There is still a trade-off between authenticity and usefulness in ITI, and the intensity of intervention needs to be adjusted according to different scenarios.

The future outlook for internal reasoning in large language and multimodal models is immensely promising. Researchers are dedicated to developing smarter and more flexible models capable of comprehending and handling complex contextual and situational nuances during the reasoning process.

Firstly, future research will focus on enhancing the models' reasoning capabilities to better understand and infer relationships among text, images, and other multimodal data. This will involve the development of more advanced model architectures and algorithms, as well as the utilization of sophisticated attention mechanisms and memory networks to capture and leverage rich contextual information.

Furthermore, a significant focus will be placed on enhancing interpretability and controllability in the models. This will empower users to better understand the models' reasoning processes and decision-making criteria, ultimately improving the efficiency and credibility of human-machine interaction and promoting the widespread application of these models in practical scenarios.

3.3.2 External Reasoning

For logical inference of problem solutions, it often necessitates reliance on antecedent's knowledge and factual underpinnings. Existing research indicates that specialized datasets are commonly utilized to assess the inferential capabilities of large models within a given field. For instance, the CSQA [265] is frequently employed in the realm of common-sense reasoning; the ScienceQA [266] is used for scientific knowledge and datasets like CommonsenseQA [265], SuperGLUE [267] are leveraged in the psychological context. While the models perform well on simple factual questions, they fall short in more complex reasoning and tasks. To improve reasoning reliability, the "Chain of Thought" prompting strategy [31] has been introduced, focusing on step-by-step reasoning. The merits of stepwise reasoning lie in its capacity to provide enhanced guidance to the LLMs in the realm of knowledge inference, consequently leading to an amelioration in the performance of the LLMs. Concurrently, empirical studies substantiate the utility and significance of the CoT in multiple intricate knowledge inference tasks [31,64,268]. Especially, ChatCoT [269], a sophisticated method to improve the problem-solving skills of LLMs. The methodology involves a dialogue-based structure that allows the LLMs to leverage external tools or their inherent reasoning capabilities in a stepwise manner. This innovative approach merges COT reasoning with the ability to manipulate tools, markedly boosting LLMs efficiency in complex tasks such as mathematics and layered question answering. The framework efficacy is underscored by its impressive performance on rigorous reasoning datasets, highlighting its potential to significantly advance chat-based LLMs in intricate reasoning tasks. Besides, Zhou et al. [236] proposed a novel NLP technique, "From Simple to Complex Prompting", which breaks down complex problems into simpler sub-problems, starting with the easiest and progressively tackling to more challenging ones. This technique provides a systematic framework for problem-solving, transforming traditional methodologies. This method is versatile, applicable to programming, scientific research, and skill acquisition.

However, due to the complexity of knowledge reasoning tasks, the performance of current LLMs still lags behind human results on tasks such as commonsense reasoning [31,270]. During the process of inference, intermediate steps are frequently disregarded, or ambiguity arises within these intermediate steps, consequently resulting in imprecise output. Fortunately, this issue can be addressed by reducing the stepwise strategy or altering the decoding method, thereby enhancing the inferential capabilities of LLMs [120,271]. For instance, Choi et al. [272] proposed KCTS, a knowledge-constrained decoding method, to address hallucinations in LLMs. Hallucination, the generation of non-factual information by LLMs, is traditionally mitigated by knowledge retrieval and model fine-tuning, but these methods are costly and risk catastrophic forgetting. KCTS overcomes these challenges by utilizing a frozen LLM, integrating a knowledge classifier and Monte-Carlo Tree Search in the decoding process, ensuring text alignment with reference knowledge. The method is model-agnostic and plug-and-play, effectively reducing hallucinations in tasks like knowledge-grounded dialogue and abstractive summarization. On changing the decoding method, Khachatryan et al. [107] examine instruction fine-tuning in language models, targeting three key areas: diversifying tasks, scaling model size and complexity, and employing chained thought data for fine-tuning. The study incorporates over 1800 tasks with various instruction templates, notably including chain thinking for complex problem formulation. Fine-tuning was conducted on models like T5, PaLM, and U-PaLM using a constant learning rate and Adafactor optimizer. Evaluations were performed on tasks such as MMLU and TyDiQA, employing both direct and chained thought prompts. Results indicate that fine-tuning based on instruction significantly enhances language comprehension and generation capabilities in the models. Moreover, the use of chained thought prompts and collaborative modes, which leverage external knowledge, markedly improved the model's performance, particularly in tasks requiring advanced reasoning and logic, thereby enhancing overall interpretability.

Furthermore, Kojima et al. [273] proposed Zero-shot-CoT reveals a surprising aspect of LLMs. Their ability to engage in zero-shot reasoning. This groundbreaking study reshapes our understanding of LLMs, overturning the conventional belief that they necessitate task-specific training. The research introduces an ingenious prompting method, proving that these models are adept at tackling intricate problems without previous training. Such an advancement heralds a new era in the application of LLMs across diverse cognitive tasks, fundamentally challenging and redefining our perceptions of their abilities and boundaries.

In the realm of LLMs and LMMs, researchers are striving to develop models that can effectively utilize external knowledge and environmental information to enhance their reasoning capabilities. Future research will focus on integrating external knowledge and environmental feedback to improve model performance. This may involve developing efficient methods for knowledge representation and retrieval, as well as designing reasoning models suitable for multimodal data. Additionally, researchers will explore utilizing feedback from the external environment to guide the model's reasoning process, potentially leveraging reinforcement learning techniques to optimize reasoning strategies. Overall, the future holds great potential for the application of external reasoning in large language and multimodal models, offering new avenues for solving complex problems and advancing AI technology.

4 Applications

In this section, we discuss the applications of LLMs in several representative fields, including medicine, finance, and other fields. The latest research in related fields shows the development potential of LLMs and LMMs in various fields.

4.1 Medicine

The medical field encompasses two integral components: healthcare and medical research, both dedicated to the enhancement of human health. Specifically, the former pertains to a domain intimately connected with everyday life, whereas the latter is typically conducted in environments such as laboratories, research institutions, and universities. Although these two areas differ, they are mutually dependent. Discoveries in medical research can guide the improvement of healthcare practices, and challenges encountered in healthcare can spark new medical research. Large models play a pivotal role in both healthcare and medical research, particularly with the advancement of artificial intelligence and machine learning technologies.

4.1.1 Healthcare

Healthcare is regarded as an indispensable component within the realm of public health. The application of large models in the healthcare sector represents an irreversible trend. The integration of large models into the medical field has been a significant achievement of contemporary research techniques. These advancements have been validated, demonstrating the multifaceted utility of large models in healthcare.

This includes, but is not limited to, conducting medical consultations [274] and conducting psychological analyses [275]. The significant application of LLMs in the medical field is exemplified by their capacity to provide crucial medical information to patients through conversational interactions [276]. This approach not only fulfills the basic healthcare needs across various social strata but also, to a certain extent, mitigates the issues of uneven distribution of medical resources and the excessive burden on the healthcare system [277]. In practical application, researchers often place heightened emphasis on the engineering of prompt words used in large models, employing specific prompting techniques to guide these models in engaging with widespread medical tasks. In a further step, researchers are pioneering the creation of LLMs that are expressly designed for the nuances and complexities of the medical industry [278–280]. For instance, Shah et al. [281] illustrated the popularity of LLMs in medicine through data. The medical LLMs based on document training and the LLMs based on medical code sequences were proposed by using medical records as training data. Another part is to point out some problems existing in current LLMs evaluation, such as unclear evaluation methods, contaminated training data sets, improper standardization checks, etc., while emphasizing the importance of correct use of LLMs and the necessity of evaluation.

From the perspective of real-time feedback, LLMs in the healthcare sector are capable of providing patients with timely health information and feedback. They offer symptom analysis and preliminary diagnoses, detailed explanations about medications and their side effects, as well as support for mental health issues. Additionally, viewed from the angle of remote communication, the application of these sophisticated language models in healthcare also serves to reduce communication costs for patients and enhance the efficiency of doctor-patient interactions [282].

Within clinical medicine, doctors often face the challenging task of sifting through an extensive amount of patient data to extract crucial information, such as allergy histories. This process, while repetitive, involves a significant workload. LLMs can assist doctors and other healthcare professionals in categorizing and swiftly extracting pertinent information from patient data, thereby streamlining this critical aspect of patient care. Certainly, the evolution of large models within the medical sector has not been without its challenges. Drawing from the unresolved issues mentioned in the preceding chapter, it becomes evident that the application of large models in the medical field also encounters challenges such as hallucinations and privacy protection [283]. These issues often precipitate grave consequences, consequently, enhancing the performance of large models stands as one of the primary concerns amongst researchers.

4.1.2 Medical Research

In the ever-evolving landscape of medicine, the introduction of LLMs marks a paradigm shift, heralding a new era of technological integration. Clusmann et al. [284] served as a pivotal starting point, outlining the broad potential and the multifaceted challenges of LLMs in this domain. This comprehensive overview sets a conceptual foundation, exploring the multifarious impacts of these advanced computational models on various aspects of medical practice and research. Building on this foundational knowledge, the focus shifts to practical applications. Study [285] delved into the innovative integration of LLMs with medical imaging, illustrating how these models can enhance diagnostic accuracy and efficiency. It symbolizes a significant leap from theoretical potential to tangible application, demonstrating the practical benefits of LLMs in enhancing the capabilities of existing medical technologies. Subsequently, the exploration of LLMs in medicine has entered the critical area of misinformation, which can be carefully addressed by building multimodal datasets [286]. This pivotal research underscores the imperative necessity to discern and mitigate the dissemination of erroneous medical information, a task made increasingly complex by the advent of sophisticated LLMs and LMMs.

Further refining the focus on LLMs utility, researchers introduced Zhongjing [287], a groundbreaking approach in enhancing Chinese medical LLMs. It utilizes a sophisticated training regime, blending diverse methodologies such as continuous pre-training, Supervised Fine-Tuning, and RLHF. Its innovation lies in its use of an extensive, real-world multi-turn medical dialogue dataset, significantly advancing the capabilities of these models in handling intricate and dynamic medical dialogues. This advancement represents a notable stride in the field of medical LLMs, especially in terms of linguistic and cultural specificity.

The narrative then explores the integration of authoritative resources with LLMs in clinical questions and answers [287]. It clearly shows that the integration of deep learning models with established medical literature significantly bolsters the capabilities of LLMs in clinical question-answering scenarios. This research illuminates a refined, integrative approach, seamlessly merging the profound, rich insights of traditional medical knowledge with the dynamic adaptability and efficiency of cutting-edge artificial intelligence technologies, thereby offering a sophisticated synthesis aimed at revolutionizing the landscape of medical diagnostics and inquiry. Furthermore, in the field of nuclear medicine, there is also a study [288] showing that the potential influence of LLMs like ChatGPT in nuclear medicine. It examines their role in enhancing medical knowledge, facilitating patient care, and aiding in research and education. The paper discusses the ability of LLMs to process vast amounts of data and provide insights, while also highlighting concerns about misinformation and data security. Emphasis is placed on the need for ethical usage, accuracy, and the critical evaluation of LLM-generated information in nuclear medicine.

LMMs in medicine to specific advancements and challenges, painting a vivid picture of a field at the cusp of a major technological revolution. This narrative illustrates the transformative power of LLMs, blending theoretical perspectives with concrete applications, and highlighting the importance of continued innovation, ethical considerations, and interdisciplinary collaboration in harnessing these powerful tools for the betterment of medical science and practice. As we stand on the brink of this new era, these articles collectively offer invaluable insights, guiding principles, and a vision for the future, where technology and medicine converge to create unprecedented opportunities for enhancing patient care, medical research, and education.

4.2 Finance

In the burgeoning field of economics, the advent of LLMs heralds a transformative era. These sophisticated models, with their unparalleled ability to analyze, interpret, and generate human-like text, are redefining traditional economic analysis and decision-making processes. By harnessing vast datasets, LLMs offer unprecedented insights into complex economic trends and behaviors, facilitating more informed and strategic economic planning. This exploration delves into how LLMs are reshaping the economic landscape, from enhancing predictive analytics to revolutionizing market research and policy formulation.

In the financial sphere, LLMs are endowed with distinctive roles and functionalities, including sophisticated financial communication [289], nuanced investment task calibration [290], and advanced emotion analysis [291]. Although most LLMs can have significant performance in the financial field, researchers will also develop large models specifically related to finance. For instance, BloombergGPT [67], a 50 billion parameter model trained on a mix of financial data and general datasets. It demonstrated that BloombergGPT significantly outperforms existing models on financial tasks without compromising performance on general LLMs benchmarks and addresses the need for domain-specific models in finance, showing the advantages of models trained on both domain-specific and general data sources; FinGPT [292], an open-source framework that provides researchers and practitioners with the tools to develop financial LLMs addresses the challenges of accessing and processing high-quality financial data.

In the field of time series analysis, a recent study [293] has adeptly harnessed the sequence modeling and interpretability of LLMs for groundbreaking financial forecasting applications. This method, involving the transformation of time series data into LLM-compatible symbolic forms and employing GPT-4 for news-derived textual summarization, has been optimized through instruction-based fine-tuning of LLaMA. This strategy, introducing multimodal financial information, bolsters prediction through inference-driven thought chains. Despite not outperforming GPT-4 Few-Shot in accuracy, its cost-effectiveness remains a strong suit. This research signals a shift in the financial time series analysis paradigm, demonstrating LLMs substantial role in deciphering intricate data. In areas like stock market forecasting, research by Xie et al. [294] has revealed the distinct prowess of large models, which address traditional methods oversight of external stochastic events. The use of tools such as ChatGPT for market trend analysis and prediction epitomizes an evolved approach to navigating the complexities of financial markets. These findings suggest that while general-purpose LLMs like ChatGPT are promising, their efficacy in specialized domains remains constrained without specific optimization, pointing to untapped potential in this arena.

4.3 Robotics

This chapter will elaborate on the application of LLMs in the field of robotics from parts human-robot interaction and external environment interaction. In addition to traditional evaluation and reasoning tasks, LLMs is also widely used in the field of robotics, such as human-robot interaction [295,296], navigation [297] and so on.

4.3.1 Human-Robot Interaction

Contemporary research delves into the utilization of LLMs in the realm of human-robot interaction, examining their prospective deployment as intricate human proxies [298]. Zhang et al. [296] proposed a review that indicates the latest advancements in the underlying structure, interface methods, and practical applications of HRI, addressing the issues of integrating LLMs into robotic systems for complex task completion. Specifically, it points out concerns and prospects of HRI in semantic understanding, ethics, privacy, and other aspects.

When it comes to specific tasks and planning, there has been an in-depth exploration of integrating LLMs with robotics technology to facilitate the convergence of task generation [299] and motion planning applications [297]. Specifically, LLM-GROP [297] is proposed for the purpose of addressing multi-object rearrangement tasks. It is a novel approach that combines commonsense knowledge derived from LLMs with a task and motion planner. This enables the translation of natural language commands into human-aligned object rearrangements in diverse environments. This signifies a step forward in robot planning methods. To test the ability of LLMs to perform actions based on behavioral instructions, the TidyBot robot [300] demonstrated a practical application of the technology, demonstrating its effectiveness in correctly classifying objects and following user-specific preferences. This method not only demonstrates a high accuracy rate in object categorization but also reflects a significant step towards more intuitive and user-friendly robotic assistants in everyday life.

In recent study, the LLM-BRAIn model [301] is adept at creating behavior trees (BTs) that are both structurally sound and logically coherent, based on textual descriptions. This capability renders it highly versatile for various robotic uses, including the operation of mobile robots and drones. To be specific, the efficiency and intuitiveness of unmanned aerial vehicle (UAV) control can be improved through the incorporation of voice and gesture interaction modalities, thereby enhancing task accuracy and user experience in UAV operations [302]. In the studied embodied system, ChatGPT has been integrated into a robot control system, an approach [303] that is observed to significantly enhance trust in human-robot interactions. This notable increase in operator trust is attributed to ChatGPT enhanced communication capabilities and its adept understanding of the nuances in human language, highlighting the critical role of advanced language processing in fostering effective human-robot collaboration.

4.3.2 External Environment Interaction

In the realm of robotics, task planning often necessitates interaction with external environments. A significant trend in this field involves the integration of human expertise with LLMs. This integration, by decoupling the planning component from machine-bound processes, simplifies the inherent complexities in planning. The results in a more adaptable method for task planning [304]. Further research has introduced an open-source platform that merges LLMs with domain expert models. This platform is designed to address complex, multi-step tasks, highlighting the versatility of LLMs in enhancing robotic task execution [305]. Additionally, leveraging prior knowledge can significantly improve a robot's performance in challenging scenarios. LLMs are instrumental in this context, facilitating improved decision-making and control adjustments, thereby boosting both the effectiveness and safety of robots during task operations [301].

In mixed reality environments, when visual information may be incomplete or misleading, integrating text information [306] can improve robot grasping type recognition in multimodal teaching to addresses the limitations of image-only methods. Key findings demonstrate that the inclusion of object affordance, derived from textual cues, significantly improves grasp-type recognition accuracy. This improvement is evident in scenarios involving both real and illusory objects. The research contributes to the field of robot teaching, offering a more effective way for robots to understand and mimic human grasping actions by combining visual and linguistic data. This advancement is particularly valuable in environments where visual information alone is insufficient, paving the way for more intuitive and efficient robot teaching methods.

Considering the multifaceted demands of real-world robotic tasks, it is inappropriate to consider task planning and motion planning alone. The work in [307] delved into the synergistic relationship between Task Planning (TP) and Motion Planning (MP) in the realm of robotics, a field collectively known as Task and Motion Planning (TAMP). It provides an extensive review of diverse algorithms, highlighting the imperative for an integrative approach that blends TP discrete decision-making with MP continuous processes. This research represents a significant stride forward in elucidating TAMP pivotal role in augmenting the functionality of robots within complex and variable environments.

4.4 Academic Research

LLMs and LMMs have broad application prospects in the field of academic research and can help researchers conduct data analysis, information processing and decision support more efficiently.

4.4.1 Science

In the realm of scientific inquiry, LLMs offer robust assistance in parsing intricate documents and distilling essential insights [308,309]. For instance, within literature reviews, GPT-4 can analyze user-uploaded texts, discerning pivotal technologies and experimental methodologies, thereby furnishing researchers with swift and precise information synopses. Moreover, LLMs can generate professionally formatted outputs, such as reports and papers, tailored to specific requirements, thus streamlining the formatting process and affording researchers more time to dedicate to their core investigative endeavors. The integration of LLMs not only amplifies research efficacy but also broadens the horizons of scientific exploration, presenting novel avenues for tackling intricate challenges.

In handling knowledge-intensive tasks, LLMs play a pivotal role [310]. Leveraging their extensive reservoir of knowledge, they swiftly and accurately retrieve relevant information, furnishing essential background knowledge and reference materials for the task at hand. Moreover, their exceptional language comprehension and generation capabilities enable them to grasp intricate contexts and produce coherent text, thereby aiding in resolving complex challenges inherent to the task. Additionally, LLMs possess the capacity to process multimodal information, amalgamating diverse modalities to enhance task efficiency. Through continuous learning and updates, these models sustain their effectiveness and applicability in knowledge-intensive tasks, furnishing robust support for seamless task execution.

Additionally, LMMs play a crucial role in the realm of scientific research, offering a versatile toolkit for analyzing and synthesizing diverse forms of data. By integrating information from multiple modalities such as text, images, and audio, these models enable researchers to gain deeper insights and perspectives into complex scientific phenomena. For example, in fields like biology and medicine, LMMs can aid in the interpretation of medical imaging data alongside textual patient records, leading to more comprehensive diagnoses and treatment plans. Similarly, in environmental science, these models can combine satellite imagery with textual weather data to better understand and predict changes in climate patterns. Overall, LMMs facilitate interdisciplinary collaboration and innovation by providing researchers with richer, more holistic datasets to draw upon in their scientific inquiries.

4.4.2 Education

LLMs can function as a supplementary teaching tool for students, aiding them in writing and reading [311,312] while also generating coherent responses across various disciplines, enhancing multidisciplinary learning. LLMs enable teachers to craft tailored plans and allocate tasks suited to individual students, thereby enhancing the efficiency and relevance of lesson preparation. Moreover, LLMs can furnish students with more comprehensive learning materials, enriching their educational experience [313,314].

Moreover, transcending the remarkable contributions of conventional LLMs to education, LMMs also offer sophisticated support within the educational domain. For the average student, LLMs employ the generation of images and videos to foster a deeper understanding of the inputted text. When it comes to teaching more abstract content, the introduction of LLMs enhances the appeal of educational methods compared to traditional approaches [315]. For students with disabilities, LMMs offer fairness and convenience in their learning and everyday lives. Through methods incorporating images, texts, and audio-visual conversion, these tools provide invaluable assistance in overcoming inherent limitations, especially for those with visual or auditory impairments. Regarding personalized learning, leveraging GPT-4 enables the customization of learning plans according to students' interests, hobbies, work-rest patterns, and other subjective factors. This approach aims to meet diverse student needs, enhance enthusiasm, and uncover potential [316].

While the widespread application of LLMs and LMMs has brought new momentum to the field of education, it is imperative not to wholly entrust students to them. One significant reason is the inherent inaccuracy often present in their outputs. Without human intervention, students may acquire erroneous knowledge, particularly within LMMs, where the complexity of generating content—such as converting text to images or images to videos—greatly amplifies the unpredictability of outputs.

Moreover, allowing students unrestricted use of these models can foster dependency, diminishing their capacity for independent thinking and impeding holistic development.

4.5 Others

Apart from the previously highlighted tasks, the deployment of LLMs and LMMs in assorted additional domains is slated for exploration. LLMs have recently found increasing application in the sophisticated areas of video and language processing, as evidenced by numerous studies such as PG-Video-LLaVA [317], PaLM-E [23], PALI [212].

In this groundbreaking model, PG-Video-LLaVA stands out as a pioneer in its field, offering pixel-level precision for enhanced understanding of video context through its audio-to-text transcription feature. The architecture seamlessly merges CLIP ViT-L/14 with an innovatively tailored visual encoder for video processing, further enhanced by a standard tracker and an advanced localization component. This component, a synergy of GroundingDINO, DEVA, and SAM, is adept at generating segmentation masks and maintaining tracking IDs. Empirical results showcase PG-Video-LLaVA exceptional spatial localization abilities, outperforming in benchmark tests like Vid-STG and HC-STVG and exceeding its peers in zero-sample visual quizzes, as seen in its use of MSRVTT-QA and MSVD-QA datasets. Notably, it excels beyond similar models, such as Video-ChatGPT, in providing detailed and contextually precise video descriptions. The mentioned studies all focus on the development of multimodal models, that is, the ability to integrate different types of inputs (such as text, images, videos) to improve processing and understanding. They are both designed to handle highly complex and diverse tasks such as robot control, visual problem solving, image description, and cross-language and cross-modal tasks. This integration of modalities enables a comprehensive approach to data processing that significantly extends the capabilities of AI systems. LLMs and LMMs show significant potential in enhancing natural resource management and ecological research [318–320]. These models excel in synthesizing diverse data sources, aiding in more accurate environmental analysis and resource estimation [321]. Furthermore, the capabilities of LMMs to handle complex visual data are underscored by several studies [322], demonstrating the effectiveness of multimodal models in performing under challenging visual conditions. This proficiency in managing diverse data types underscores the broader applicability of these models. Additionally, LLMs and LMMs have also demonstrated their utility in enhancing geospatial and semantic analyses across various domains [323–325].

In summary, the application of LMMs is seen as highly promising for enhancing fairness. Through the integration of visual, textual, and other data forms, an accurate response to diverse requirements can be achieved. Mitigating biases inherent in single-source data improves the fairness and inclusivity of services. For instance, in fields such as healthcare and education, data can be analyzed more comprehensively. It ensures that equitable services are provided to users from varied backgrounds. To be specific, traditional LLMs architectures need to be improved to accommodate multi-modal data processing. For example, the fusion of the ViT [96] and the BERT [12] model achieves a unified architecture for processing visual and text data. As LMMs mature, application scenarios have gradually expanded. From initial image description generation, visual question answering, to complex cross-modal retrieval and augmented reality applications, LMMs have demonstrated their powerful ability to fuse and understand multiple sources of information. The transformation from large language models to multi-modal models marks the advancement of AI from single-modal processing to multi-modal fusion. In this process, technological breakthroughs such as expanded data input types, cross-modal embedding, improved model architecture, and joint training strategies have enabled AI systems to demonstrate stronger understanding and generation capabilities in complex and diverse application scenarios. This transformation not only increases the breadth and depth of AI applications, but also provides a broader development space for future intelligent systems.

5 Discussion

LLMs and LMMs continue to hold vast potential for future development. Establishing comprehensive and challenging cross-modal datasets and benchmarks is a crucial direction for the future. This will enable researchers to assess and compare the performance of different models, facilitating their application in more complex scenarios. In multimodal models, effectively integrating and understanding information from various senses, such as vision, hearing, and touch, is vital for enhancing natural human-machine interaction. Moreover, as the scale of models increases, so does their energy consumption and environmental impact. Future research should explore designing more energy-efficient model training processes, such as by improving algorithm efficiency, optimizing hardware usage, or employing green energy sources. Additionally, cultivating an open, collaborative research environment will accelerate technological advances and promote the widespread adoption of large model technologies. The ultimate goal is to create models that not only excel in performance but also operate transparently and ethically, making a positive impact on society.

6 Conclusion

In this paper, we explored the transition from LLMs to LMMs, emphasizing the development and integration of AI systems capable of understanding various data formats beyond text. We introduced the foundational concepts of attention in LLMs and LMMs, explored the structure and architectures of both LLMs and LMMs, discussed training methods and data sources, and examined the emergent abilities of LLMs, including ICL, instruction following and step-by-step reasoning. We discussed the unresolved issues confronting large language models and multi-modal models. These issues include context understanding, illusion correction, cognitive ability assessment, and accuracy reasoning. Additionally, we presented new research findings in various fields. We highlighted the specific applications of large language models and multimodal models in various fields, including medicine, economics, robotics and others. Furthermore, we discussed the potential for these models to be utilized across different domains in the future. This paper summarized technological advancements, potential applications, and challenges related to data integration, cross-modal comprehension, and ethical considerations.

Acknowledgement: Thank all the people who contributed to this paper. Stand on the shoulders of giants, in memory of Galileo Galilei (1564–1642).

Funding Statement: We acknowledge funding from NSFC Grant 62306283.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Zheyi Chen, Liuchang Xu; data collection: Hongting Zheng; analysis and interpretation of results: Luyao Chen, Keping Yu; draft manuscript preparation: Liang Zhao, Amr Tolba, Hailin Feng. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials used in this review are derived from publicly accessible databases and previously published studies, which are cited throughout the text. References to these sources are provided in the bibliography.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Naveed *et al.*, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[2] S. Pinker, "ChatterBoxes," in *The Language Instinct: How the Mind Creates Language*. UK: Penguin, vol. 1, 2003, pp. 34–45.

[3] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language: What is it, who has it, and how did it evolve?" *Science*, vol. 298, no. 5598, pp. 1569–1579, 2002. doi: 10.1126/science.298.5598.1569.

[4] A. M. Turing, "Learning machines," in *Computing Machinery and Intelligence*. UK: Springer, Oxford, 2009, pp. 437–439.

[5] W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[6] F. Jelinek, *Statistical Methods for Speech Recognition*. MIT press, Cambridge, Massachusetts, USA, 1998, pp. 297–298.

[7] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proc. IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000. doi: 10.1109/5.880083.

[8] N. Ide and J. Véronis, "Introduction to the special issue on word sense disambiguation: The state of the art," *Comput. Linguist.*, vol. 24, no. 1, pp. 1–40, 1998.

[9] A. Stolcke, "SRILM–an extensible language modeling toolkit," in *Proc. 7th Int. Conf. Spoken Lang. Process. (ICSLP 2002)*, Denver, CO, USA, 2002, pp. 901–904. doi: 10.21437/ICSLP.2002-303.

[10] S. M. Thede and M. Harper, "A second-order hidden Markov model for part-of-speech tagging," in *Proc. 37th Annu. Meet. Assoc. Comput. Linguist.*, University of Maryland, MD, USA, 1999, pp. 175–182.

[11] C. Zhai, "Statistical language models for information retrieval a critical review," *Found. Trends® Inf. Retr.*, vol. 2, no. 3, pp. 137–213, 2008. doi: 10.1561/1500000008.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[14] F. García-Peñalvo and A. Vázquez-Ingelmo, "What do we mean by GenAI? a systematic mapping of the evolution, trends, and techniques involved in generative AI," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, pp. 7–16, 2023.

[15] J. B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 23716–23736.

[16] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[17] K. Li *et al.*, "VideoChat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[18] 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.

[19] X. Mei *et al.*, "WavCaps: A ChatGPT-assisted weaklylabelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[20] C. Lyu *et al.*, "LLM: Multi-modal language modeling with image, audio, video, and text integration," *arXiv preprint arXiv:2306.09093*, 2023.

[21] S. Huang *et al.*, "Language is not all you need: Aligning perception with language models," *arXiv preprint arXiv:2302.14045*, 2023.

[22] Y. Cao *et al.*, "A comprehensive survey of AI-generated content (AIGC): A history of generative ai from GAN to ChatGPT," *arXiv preprint arXiv:2303.04226*, 2023.

[23] D. Driess *et al.*, "PaLM-E: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[24] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual ChatGPT: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023.

[25] J. Achiam *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[26] Y. Wang *et al.*, "Aligning large language models with human: A survey," *arXiv preprint arXiv:2307.12966*, 2023.

[27] T. Brown *et al.*, "Language models are few-shot learners," *Adv. Neural Inf. Process Syst.*, vol. 33, pp. 1877–1901, 2020.

[28] A. Chowdhery *et al.*, "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, pp. 1–113, 2023.

[29] R. Taylor *et al.*, "Galactica: A large language model for science," *arXiv preprint arXiv:2211.09085*, 2022.

[30] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[31] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24824–24837, 2022.

[32] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process Syst.*, vol. 30, pp. 5998–6008, 2017.

[33] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[34] N. Shazeer, "Fast transformer decoding: One write-head is all you need," *arXiv preprint arXiv:1911.02150*, 2019.

[35] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 16344–16359, 2022.

[36] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 8026–8037, 2019.

[37] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2020, pp. 3505–3506.

[38] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper and B. Catanzaro, "Megatron-LM: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[39] T. Dao, "FlAshattention-2: Faster attention with better parallelism and work partitioning," *arXiv preprint arXiv:2307.08691*, 2023.

[40] W. Kwon *et al.*, "Efficient memory management for large language model serving with pagedattention," in *Proc. 29th Symp. Oper. Syst. Princ.*, Koblenz, Germany, 2023, pp. 611–626.

[41] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image– text matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5412–5425, 2020. doi: 10.1109/TNNLS.2020.2967597.

[42] J. Cai *et al.*, "A novel graph-attention based multimodal fusion network for joint classification of hyperspectral image and lidar data," *Expert. Syst. Appl.*, vol. 249, no. 3, pp. 123587, 2024. doi: 10.1016/j.eswa.2024.123587.

[43] V. S. Dorbala, G. Sigurdsson, R. Piramuthu, J. Thomason, and G. S. Sukhatme, "CLIP-Nav: Using CLIP for zero-shot vision-andlanguage navigation," *arXiv preprint arXiv:2211.16649*, 2022.

[44] S. A. Sontakke *et al.*, "RoboCLIP: One demonstration is enough to learn robot policies," *arXiv preprint arXiv:2310.07899*, 2023.

[45] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2024, pp. 26296–26306.

[46] Y. Sun *et al.*, "ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[47] W. Zeng *et al.*, "PanGu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation," *arXiv preprint arXiv:2104.12369*, 2021.

[48] Z. Zhang *et al.*, "CPM-2: Large-scale cost-effective pre-trained language models," *AI Open*, vol. 2, no. 1, pp. 216–224, 2021. doi: 10.1016/j.aiopen.2021.12.003.

[49] S. Wang *et al.*, "ERNIE 3.0 Titan: Exploring large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2112.12731*, 2021.

[50] S. Black *et al.*, "GPT-NeoX-20B: An open-source autoregressive language model," *arXiv preprint arXiv:2204.06745*, 2022.

[51] B. Workshop *et al.*, "BLOOM: A 176B-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[52] N. Du *et al.*, "GLaM: Efficient scaling of language models with mixture-of-experts," in *Int. Conf. Mach. Learn.*, Baltimore, MD, USA, 2022, pp. 5547–5569.

[53] R. Thoppilan *et al.*, "LaMDA: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[54] Y. Tay *et al.*, "UL2: Unifying language learning paradigms," in *Int. Conf. Learn. Represent.*, 2022.

[55] A. Zeng *et al.*, "GLM-130B: An open bilingual pre-trained model," *arXiv preprint arXiv:2210.02414*, 2022.

[56] O. Lieber, O. Sharir, B. Lenz, and Y. Shoham, "Jurassic-1: Technical details and evaluation," in *White Paper*, *AI21 Labs*, vol. 1, pp. 9, 2021.

[57] B. Kim *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.

[58] S. Wu *et al.*, "Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning," *arXiv preprint arXiv:2110.04725*, 2021.

[59] X. Ren *et al.*, "PanGu-Σ: Towards trillion parameter language model with sparse heterogeneous computing," *arXiv preprint arXiv:2303.10845*, 2023.

[60] X. Zhang and Q. Yang, "XinYuan 2.0: A large chinese financial chat model with hundreds of billions parameters," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manag.*, Birmingham, UK, 2023, pp. 4435–4439.

[61] S. Zhang *et al.*, "OPT: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[62] J. W. Rae *et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint arXiv:2112.11446*, 2021.

[63] J. Hoffmann *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[64] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Trans. Assoc. Comput. Linguist.*, vol. 7, pp. 249–266, 2019. doi: 10.1162/tacl_a_00266.

[65] Y. Tay *et al.*, "Transcending scaling laws with 0.1% extra compute," *arXiv preprint arXiv:2210.11399*, 2022.

[66] S. Soltan *et al.*, "AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model," *arXiv preprint arXiv:2208.01448*, 2022.

[67] S. Wu *et al.*, "BloombergGPT: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[68] L. Xue *et al.*, "mT5: A massively multilingual pretrained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, pp. 11934, 2010.

[69] Y. Li *et al.*, "Competitionlevel code generation with AlphaCode," *Science*, vol. 378, no. 6624, pp. 1092–1097, 2022. doi: 10.1126/science.abq1158.

[70] E. Nijkamp *et al.*, "CodeGen: An open large language model for code with multi-turn program synthesis," *arXiv preprint arXiv:2203.13474*, 2022.

[71] 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.

[72] K. Q. Zhou and H. Nabus, "The ethical implications of DALL-E: Opportunities and challenges," *Mesopo. J. Comput. Sci.*, vol. 2023, pp. 16–21, 2023. doi: 10.58496/MJCSC/2023/003.

[73] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, pp. 3, 2022.

[74] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36479–36494, 2022.

[75] J. Yu *et al.*, "Scaling autoregressive models for content-rich text-to-image generation," *arXiv preprint arXiv:2206.10789*, vol. 2, no. 3, pp. 5, 2022.

[76] 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.

[77] C. Jia *et al.*, "Scaling up visual and visionlanguage representation learning with noisy text supervision," in *Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.

[78] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[79] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[80] W. Wang *et al.*, "CogVLM: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[81] R. Anil *et al.*, "PaLM 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[82] Z. Luo *et al.*, "WizardCoder: Empowering code large language models with evol-instruct," *arXiv preprint arXiv:2306.08568*, 2023.

[83] J. Bai *et al.*, "Qwen-VL: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[84] B. Lin *et al.*, "MoE-LLaVA: Mixture of experts for large vision-language models," *arXiv preprint arXiv:2401.15947*, 2024.

[85] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 3104–3108, 2014.

[86] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[87] P. J. Liu *et al.*, "Generating wikipedia by summarizing long sequences," *arXiv preprint arXiv:1801.10198*, 2018.

[88] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[89] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[90] 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.

[91] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process Syst.*, vol. 32, pp. 5753–5763, 2019.

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

[93] K. Clark, M. -T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pretraining text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[94] D. Zhang *et al.*, "MM-LLMs: Recent advances in multimodal large language models," *arXiv preprint arXiv:2401.13601*, 2024.

[95] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, vol. 139, pp. 1059–1071.

[96] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021.

[97] Y. Fang *et al.*, "Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 19358–19369.

[98] P. Zhou *et al.*, "A survey on generative ai and LLM for video generation, understanding, and streaming," *arXiv preprint arXiv:2404.16038*, 2024.

[99] F. Chen *et al.*, "X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages," *arXiv preprint arXiv:2305.04160*, 2023.

[100] W. -N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[101] S. Chen *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[102] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Int. Conf. Mach. Learn.*, Honolulu, HI, USA, 2023, pp. 28492–28518.

[103] H. W. Chung *et al.*, "Scaling instructionfinetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[104] W. -L. Chiang *et al.*, "Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality," vol. 2, no. 3, pp. 6. 2023. Accessed: Aug. 14, 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[105] M. Zhao, F. Bao, C. Li, and J. Zhu, "EGSDE: Unpaired image-to-image translation via energy-guided stochastic differential equations," *Adv. Neural Inf. Process Syst.*, vol. 35, pp. 3609–3623, 2022.

[106] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 10684–10695.

[107] L. Khachatryan *et al.*, "Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Paris, France, 2023, pp. 15954–15964.

[108] H. Liu *et al.*, "AudioLDM: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[109] I. Hou, O. Man, S. Mettille, S. Gutierrez, K. Angelikas, and S. MacNeil, "More robots are coming: Large multimodal models (ChatGPT) can solve visually diverse images of parsons problems," in *Proc. 26th Australas. Comput. Educ. Conf.*, Sydney, NSW, Australia, 2024, pp. 29–38.

[110] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 27730–27744, 2022.

[111] S. Iyer *et al.*, "OPT-IML: Scaling language model instruction meta learning through the lens of generalization," *arXiv preprint arXiv:2212.12017*, 2022.

[112] Q. Ye *et al.*, "mPLUG-Owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[113] W. Dai *et al.*, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[114] W. Wang *et al.*, "VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks," *arXiv preprint arXiv:2305.11175*, 2023.

[115] Z. Xu, Y. Shen, and L. Huang, "Multiinstruct: Improving multimodal zero-shot learning via instruction tuning," *arXiv preprint arXiv:2212.10773*, 2022.

[116] S. Yin *et al.*, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[117] T. Gupta, A. Kamath, A. Kembhavi, and D. Hoiem, "Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 16399–16409.

[118] Z. Zhao *et al.*, "ChatBridge: Bridging modalities with large language model as a language catalyst," *arXiv preprint arXiv:2305.16103*, 2023.

[119] L. Li *et al.*, "M^{3}IT: A large-scale dataset towards multi-modal multilingual instruction tuning," *arXiv preprint arXiv:2306.04387*, 2023.

[120] Y. Wang *et al.*, "Self-instruct: Aligning language model with self-generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.

[121] R. Yang *et al.*, "GPT4Tools: Teaching large language model to use tools via selfinstruction," *arXiv preprint arXiv:2305.18752*, 2023.

[122] R. Pi *et al.*, "DetGPT: Detect what you need via reasoning," *arXiv preprint arXiv:2305.14167*, 2023.

[123] G. Luo, Y. Zhou, T. Ren, S. Chen, X. Sun and R. Ji, "Cheap and quick: Efficient vision-language instruction tuning for large language models," *arXiv preprint arXiv:2305.15023*, 2023.

[124] R. Zhang *et al.*, "LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

[125] J. Wei *et al.*, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[126] J. Ge, H. Luo, S. Qian, Y. Gan, J. Fu and S. Zhan, "Chain of thought prompt tuning in vision language models," *arXiv preprint arXiv:2304.07919*, 2023.

[127] Z. Yang *et al.*, "MM-REACT: Prompting ChatGPTfor multimodal reasoning and action," *arXiv preprint arXiv:2303.11381*, 2023.

[128] Y. Zhu *et al.*, "Aligning books and movies: Towards storylike visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 19–27.

[129] M.S. Hart *et al.*, "Project Gutenberg," Accessed: Jan. 06, 2024. [Online]. Available: https://www.gutenberg.org/

[130] Y. Bai *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[131] F. Dolly *et al.*, "Introducing the world's first truly open instruction-tuned LLM," *Databricks*, 2023. [Online]. Accessed: Apr. 04, 2024. Available: https://simonwillison.net/2023/Apr/13/dolly/

[132] B. Guo *et al.*, "How close is ChatGPT to human experts? comparison corpus, evaluation, and detection," *arXiv preprint arXiv:2301.07597*, 2023.

[133] A. Köpf *et al.*, "Open Assistant conversations-democratizing large language model alignment," *arXiv preprint arXiv.2304.07327*, 2023.

[134] S. Tey *et al.*, "ShareGPT," 2023. Accessed: Jan. 06, 2023. [Online]. Available: https://sharegpt.com/

[135] E. Bisong, "Google BigQuery," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, Berkeley, CA: Apress, 2019, pp. 485–517.

[136] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," *arXiv preprint arXiv:1806.02847*, 2018.

[137] R. Zellers *et al.*, "Defending against neural fake news," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 9054–9065, 2019.

[138] T. Tang, J. Li, W. X. Zhao, and J. R. Wen, "MVP: Multi-task supervised pre-training for natural language generation," *arXiv preprint arXiv:2206.12131*, 2022.

[139] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, "Cross-task generalization via natural language crowdsourcing instructions," *arXiv preprint arXiv:2104.08773*, 2021.

[140] C. Schuhmann *et al.*, "The OIG dataset," 2023. Accessed: Apr. 07, 2024. [Online]. Available: https://laion.ai/blog/oig-dataset/

[141] S. H. Bach *et al.*, "PromptSource: An integrated development environment and repository for natural language prompts," *arXiv preprint arXiv:2202.01279*, 2022.

[142] Y. Wang *et al.*, "Super-Naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks," *arXiv preprint arXiv:2204.07705*, 2022.

[143] N. Muennighoff *et al.*, "Crosslingual generalization through multitask finetuning," *arXiv preprint arXiv:2211.01786*, 2022.

[144] A. Gokalan *et al.*, "OpenWebText corpus," 2023. Accessed: Apr. 07, 2024. [Online]. Available: https://skylion007.github.io/OpenWebTextCorpus/

[145] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proc. Int. AAAI Conf. Web Social Media*, 2020, vol. 14, pp. 830–839. doi: 10.1609/icwsm.v14i1.7347.

[146] R. Taori *et al.*, "Stanford Alpaca: An instruction-following Llama model," vol. 1, no. 9,2023.Accessed: Apr. 07, 2024.https://github.com/tatsu-lab/stanford_alpaca/

[147] C. Xu, D. Guo, N. Duan, and J. McAuley, "Baize: An open-source chat model with parameter-efficient tuning on self-chat data," *arXiv preprint arXiv:2304.01196*, 2023.

[148] Y. Ji *et al.*, "Towards better instruction following language models for Chinese: Investigating the impact of training data and evaluation," *arXiv preprint arXiv:2304.07854*, 2023.

[149] C. Josephus, "Guanaco-generative universal assistant for natural-language adaptive context-aware omnilingual outputs.," 2023. Accessed: Apr. 07, 2024. [Online]. Available: https://guanaco-model.github.io/

[150] G. Penedo *et al.*, "The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only," *arXiv preprint arXiv:2306.01116*, 2023.

[151] J. Wales *et al.*, "I. positive and wikipedia," 2001. Accessed: Apr. 06, 2024. [Online]. Available: https://en.wikipedia.org/wiki/

[152] M. Weber *et al.*, "RedPajama-data: An open-source recipe to reproduce llama training dataset," 2023. Accessed: Apr. 04, 2024. [Online]. Available: https://github.com/open-models-platform/openmodels.RedPajama-Data/

[153] H. Laurençon *et al.*, "The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 31809–31826, 2022.

[154] L. Gao *et al.*, "The Pile: An 800GB dataset of diverse text for language modeling," *arXiv preprint arXiv:2101.00027*, 2020.

[155] D. Kocetkov *et al.*, "The Stack: 3 TB of permissively licensed source code," *arXiv preprint arXiv:2211.15533*, 2022.

[156] C. Zhou *et al.*, "LIMA: Less is more for alignment," *arXiv preprint arXiv:2305.11206*, 2023.

[157] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[158] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.

[159] D. Ganguli *et al.*, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *arXiv preprint arXiv:2209.07858*, 2022.

[160] G. Xu *et al.*, "CValues: Measuring the values of chinese large language models from safety to responsibility," *arXiv preprint arXiv:2307.09705*, 2023.

[161] J. Dai *et al.*, "Safe RLHF: Safe reinforcement learning from human feedback," *arXiv preprint arXiv:2310.12773*, 2023.

[162] R. Liu *et al.*, "Training socially aligned language models in simulated human society," *arXiv preprint arXiv:2305.16960*, 2023.

[163] K. Ethayarajh, Y. Choi, and S. Swayamdipta, "Understanding dataset difficulty with V-usable information," in *Int. Conf. Mach. Learn.*, Baltimore, Maryland, USA, 2022, pp. 5988–6008.

[164] N. Lambert, L. Tunstall, N. Rajani, and T. Thrush, "HuggingFace H4 stack exchange preference dataset," 2023. Accessed: Apr. 04, 2024. [Online]. Available: https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences/

[165] N. Stiennon *et al.*, "Learning to summarize with human feedback," *Adv. Neural Inf. Process Syst.*, vol. 33, pp. 3008–3021, 2020.

[166] R. Nakano *et al.*, "WebGPT: Browser-assisted question-answering with human feedback," *arXiv preprint arXiv:2112.09332*, 2021.

[167] M. Byeon, B. Park, H. Kim, S. Lee, W. Baek and S. Kim, "COYO-700M: Image-text pair dataset," 2022. Accessed: Apr. 04, 2024. [Online]. Available: https://github.com/kakaobrain/coyo-dataset/

[168] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3558–3568.

[169] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image ALT-text dataset for automatic image captioning," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist.*, Melbourne, Australia, 2018, pp. 2556–2565.

[170] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Veneto, Italy, 2017, pp. 843–852.

[171] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, 2022, pp. 12104–12113.

[172] T. -Y. Lin, "Microsoft COCO: Common objects in context," in *Comput. Vis.–ECCV 2014: 13th Eur. Conf.*, Zurich, Switzerland, Springer, 2014, pp. 740–755.

[173] X. Chen *et al.*, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[174] D. A. Hudson and C. D. Manning, "GQA: A new dataset for realworld visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 6700–6709.

[175] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "OCR-VQA: Visual question answering by reading text in images," in *2019 Int. Conf. Document Anal. Recognit. (ICDAR)*, International Convention Centre (ICC) Sydney, 2019, pp. 947–952.

[176] U. Naseem, M. Khushi, A. G. Dunn, and J. Kim, "K-PathVQA: Knowledge-aware multimodal representation for pathology visual question answering," *IEEE J. Biomed. Health Inform.*, vol. 28, pp. 1886–1895, 2023.

[177] E. Choi *et al.*, "QuAC: Question answering in context," *arXiv preprint arXiv:1808.07036*, 2018.

[178] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. 2014 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 787–798.

[179] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Comput. Vis.-ECCV 2016: 14th Euro. Conf.*, Amsterdam, The Netherlands, Springer, 2016, pp. 69–85.

[180] B. Liu, L. M. Zhan, L. Xu, L. Ma, Y. Yang and X. M. Wu, "Slake: A semantically-labeled knowledgeenhanced dataset for medical visual question answering," in *2021 IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Washington, DC, USA, 2021, pp. 1650–1654.

[181] A. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha, "LaTr: Layout-aware transformer for scene-text VQA," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 16548–16558.

[182] A. Singh *et al.*, "Towards VQA models that can read," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 8317–8326.

[183] R. Krishna *et al.*, "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017. doi: 10.1007/s11263-016-0981-7.

[184] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Visi. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4995–5004.

[185] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Sci. Data*, vol. 5, no. 1, pp. 1–10, 2018. doi: 10.1038/sdata.2018.251.

[186] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6904–6913.

[187] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-OKVQA: A benchmark for visual question answering using world knowledge," in *Euro. Conf. Comput. Vis.*, Tel Aviv, Israel, Springer, 2022, pp. 146–162.

[188] S. Yitzhak Gadre *et al.*, "DataComp: In search of the next generation of multimodal datasets," *arXiv preprint arXiv:2304.14108*, 2023.

[189] M. Mathew, D. Karatzas, and C. Jawahar, "DocVQA: A dataset for VQA on document images," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, 2021, pp. 2200–2209.

[190] K. Kafle, B. Price, S. Cohen, and C. Kanan, "DVQA: Understanding data visualizations via question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 5648–5656.

[191] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, "RedCaps: Webcurated image-text data created by the people," *arXiv preprint arXiv:2111.11431*, 2021.

[192] V. Ordonez, G. Kulkarni, and T. Berg, "Im2Text: Describing images using 1 million captioned photographs," *Adv. Neural Inf. Process. Syst.*, vol. 24, pp. 1143–1151, 2011.

[193] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proc. 44th Int. ACM SIGIR Con. Res. Develop. Inform. Retr.*, 2021, pp. 2443–2449.

[194] B. Thomee *et al.*, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016. doi: 10.1145/2812802.

[195] J. Wu *et al.*, "AI challenger: A large-scale dataset for going deeper in image understanding," *arXiv preprint arXiv:1711.06475*, 2017.

[196] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Computat. Linguist.*, vol. 2, no. 1, pp. 67–78, 2014. doi: 10.1162/tacl_a_00166.

[197] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier and S. Lazebnik, "Flickr30k entities: Collecting regionto-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 2641–2649.

[198] M. Singh *et al.*, "Revisiting weakly supervised pre-training of visual perception models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 804–814.

[199] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017. doi: 10.1145/3065386.

[200] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[201] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[202] C. Schuhmann *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," *Adv. in Neural Inf. Process. Syst.*, vol. 35, pp. 25278–25294, 2022.

[203] C. Schuhmann *et al.*, "LAION-400M: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[204] C. Schuhmann, A. Köpf, R. Vencu, T. Coombes, and R. Beaumont, "LAION COCO: 600m synthetic captions from LAION 2B-EN," Accessed: Apr. 07, 2024. [Online]. Available: https://laion.ai/blog/laion-coco/

[205] H. Agrawal *et al.*, "nocaps: Novel object captioning at scale," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Republic of Korea, 2019, pp. 8948–8957.

[206] B. -L Lu and W. -L Schuhmann, "Seed dataset," Accessed: Apr. 07, 2024. [Online]. Available: https://bcmi.sjtu.edu.cn/home/seed/

[207] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 1728–1738.

[208] P. Lu *et al.*, "Learn to explain: Multimodal reasoning via thought chains for science question answering," 2022. Accessed: Apr. 07, 2024. [Online]. Available: https://scienceqa.github.io/

[209] H. Laurençon *et al.*, "OBELISC: An open web-scale filtered dataset of interleaved image-text documents," *arXiv preprint arXiv:2306.16527*, 2023.

[210] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 5288–5296.

[211] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "TextCaps: A dataset for image captioning with reading comprehension," in *Comput. Vis.–ECCV 2020: 16th Euro. Conf.,* Glasgow, UK, Online, Springer, 2020, pp. 742–758.

[212] X. Chen *et al.*, "PaLI: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[213] J. Gu *et al.*, "Wukong: A 100 million largescale Chinese cross-modal pre-training benchmark," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 26418–26431, 2022.

[214] X. Chen *et al.*, "PaLI-X: On scaling up a multilingual vision and language model," *arXiv preprint arXiv:2305.18565*, 2023.

[215] W. Zhu *et al.*, "Multimodal C4: An open, billion-scale corpus of images interleaved with text," *arXiv preprint ArXiv:2304.06939*, 2023.

[216] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming mandarin ASR research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[217] W. Wu *et al.*, "MOFI: Learning image representations from noisy entity annotated images," *arXiv preprint arXiv:2306.07952*, 2023.

[218] J. Dodge *et al.*, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," *arXiv preprint arXiv:2104.08758*, 2021.

[219] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, Silver Spring, MD, USA, 2009, pp. 248–255.

[220] G. Team *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[221] D. Hendrycks *et al.*, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[222] Y. Huang *et al.*, "C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models," *arXiv preprint arXiv:2305.08322*, 2023.

[223] K. Cobbe *et al.*, "Training verifiers to solve math word problems," *arXiv preprintarXiv: 2110. 14168*, 2021.

[224] M.Suzgun *et al.*, "Challenging big-bench tasks and whether chain-of-thought can solve them," *arXiv preprint arXiv:2210.09261*, 2022.

[225] D. Hendrycks *et al.*, "Measuring coding challenge competence with apps," *arXiv preprint arXiv:2105.09938*, 2021.

[226] C. Fu *et al.*, "Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis," *arXiv preprint arXiv:2405.21075*, 2024.

[227] Z. Yin *et al.*, "LAMM: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark," *arXiv preprint arXiv:2306.06687*, 2023.

[228] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-ChatGPT: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[229] P. Liang *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2306.05685*, 2022.

[230] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-bench and chatbot arena," *arXiv preprint arXiv:2306.05685*, 2023.

[231] Y. Dubois *et al.*, "AlpacaFarm: A simulation framework for methods that learn from human feedback," 2024, Accessed: Apr. 06, 2024. [Online]. Available: https://arxiv.org/abs/2305.14387/

[232] Y. Zhao, M. Khalman, R. Joshi, S. Narayan, M. Saleh and P. J. Liu, "Calibrating sequence likelihood improves conditional language generation," *arXiv preprint arXiv:2210.00045*, 2022.

[233] J. Wei *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[234] B. A. Huberman and T. Hogg, "Phase transitions in artificial intelligence systems," *Artif. Intell.*, vol. 33, no. 2, pp. 155–171, 1987. doi: 10.1016/0004-3702(87)90033-6.

[235] V. Sanh *et al.*, "Multitask prompted training enables zero-shot task generalization," *arXiv preprint arXiv:2110.08207*, 2021.

[236] D. Zhou *et al.*, "Least-to-most prompting enables complex reasoning in large language models," *arXiv preprint arXiv:2205.10625*, 2022.

[237] Y. Fu, H. Peng, and T. Khot, "How does GPT obtain its ability? Tracing emergent abilities of language models to their sources," *Yao Fu's Notion*, 2022. Accessed: Apr. 07, 2024. [Online]. Available: https://franxyao.github.io/

[238] 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.

[239] Y. S. Chuang, Y. Xie, H. Luo, Y. Kim, J. Glass and P. He, "DoLa: Decoding by contrasting layers improves factuality in large language models," *arXiv preprint arXiv:2309.03883*, 2023.

[240] G. Izacard *et al.*, "Few-shot learning with retrieval augmented language models," *arXiv preprint arXiv:2208.03299*, 2022.

[241] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Int. Conf. Mach. Learn.*, 2020, pp. 3929–3938.

[242] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

[243] Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao, and J. R. Wen, "Complex knowledge base question answering: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 11, pp. 11196–11215, 2022. doi: 10.1109/TKDE.2022.3223858.

[244] B. Peng *et al.*, "Check your facts and try again: Improving large language models with external knowledge and automated feedback," *arXiv preprint arXiv:2302.12813*, 2023.

[245] S. Qi, Z. Cao, J. Rao, L. Wang, J. Xiao and X. Wang, "What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing," *Inform. Process. Manage.*, vol. 60, no. 6, pp. 103510, 2023. doi: 10.1016/j.ipm.2023.103510.

[246] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 19113–19122.

[247] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confid. Comput.*, pp. 100211, 2024.

[248] C. Mitra, B. Huang, T. Darrell, and R. Herzig, "Compositional chain-of-thought prompting for large multimodal models," *arXiv preprint arXiv:2311.17076*, 2023.

[249] R. Cai *et al.*, "BenchLMM: Benchmarking cross-style visual capability of large multimodal models," *arXiv preprint arXiv:2312.02896*, 2023.

[250] Z. Ren *et al.*, "PixeLLM: Pixel reasoning with large multimodal model," *arXiv preprint arXiv:2312.02228*, 2023.

[251] Z. Liu *et al.*, "Deja Vu: Contextual sparsity for efficient LLMs at inference time," in *Int. Conf. Mach. Learn.*, PMLR, Honolulu, HI, USA, 2023, pp. 22137–22176.

[252] P. Manakul, A. Liusie, and M. J. Gales, "SelfcheckGPT: Zeroresource black-box hallucination detection for generative large language models," *arXiv preprint arXiv:2303.08896*, 2023.

[253] R. Friel and A. Sanyal, "Chainpoll: A high efficacy method for LLM hallucination detection," *arXiv preprint arXiv:2310.18344*, 2023.

[254] J. Fu, S. K. Ng, Z. Jiang, and P. Liu, "GPTscore: Evaluate as you desire," *arXiv preprint arXiv:2302. 04166*, 2023.

[255] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, May 2023.

[256] O. Honovich, A. Hassidim, and Y. Matias, "True: Re-evaluating factual consistency evaluation," *arXiv preprint arXiv:2204.04991*, 2022.

[257] O. B. Mercea, L. Riesch, A. Koepke, and Z. Akata, "Audio-visual generalised zero-shot learning with cross-modal attention and language," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 10553–10563.

[258] Y. Liu *et al.*, "Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 3042–3051.

[259] Y. L. Sung, J. Cho, and M. Bansal, "VL-adapter: Parameter efficient transfer learning for vision-and-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 5227–5237.

[260] Z. Sun *et al.*, "Aligning large multimodal models with factually augmented RLHF," *arXiv preprint arXiv:2309.14525*, 2023.

[261] Y. Lu, C. Li, H. Liu, J. Yang, J. Gao and Y. Shen, "An empirical study of scaling instruct-tuned large multimodal models," *arXiv preprint arXiv:2309.09958*, 2023.

[262] K. H. Huang, H. P. Chan, and H. Ji, “Zero-shot faithful factual error correction,” *arXiv preprint arXiv:2305.07982*, 2023.
[263] L. K. Umapathi, A. Pal, and M. Sankarasubbu, “Med-HALT: Medical domain hallucination test for large language models,” *arXiv preprint arXiv:2307.15343*, 2023.
[264] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg, “Inference-time intervention: Eliciting truthful answers from a language model,” *arXiv preprint arXiv:2306.03341*, 2023.
[265] A. Talmor, J. Herzig, N. Lourie, and J. Berant, “CommonsenseQA: A question answering challenge targeting commonsense knowledge,” *arXiv preprint arXiv:1811.00937*, 2018.
[266] T. Saikh, T. Ghosal, A. Mittal, A. Ekbal, and P. Bhattacharyya, “ScienceQA: A novel resource for question answering on scholarly articles,” *Int. J. Digit. Libr.*, vol. 23, no. 3, pp. 289–301, 2022. doi: 10.1007/s00799-022-00329-y.
[267] P. -E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperGlue: Learning feature matching with graph neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4938–4947.
[268] 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.
[269] Z. Chen, K. Zhou, B. Zhang, Z. Gong, W. X. Zhao and J. R. Wen, “ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models,” *arXiv preprint arXiv:2305.14323*, 2023.
[270] S. Dhingra, M. Singh, S. B. Vaisakh, N. Malviya, and S. Singh Gill, “Mind meets machine: Unravelling GPT-4’s cognitive psychology,” *arXiv preprint arXiv:2303.11436*, 2023.
[271] Y. Li *et al.*, “On the advance of making language models better reasoners,” *arXiv preprint arXiv:2206.02336*, 2022.
[272] S. Choi, T. Fang, Z. Wang, and Y. Song, “KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection,” *arXiv preprint arXiv:2310.09044*, 2023.
[273] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 22199–22213, 2022.
[274] O. Nov, N. Singh, and D. M. Mann, “Putting ChatGPT’s medical advice to the (turing) test,” *JMIR Med. Educ.*, vol. 10, no. 9, pp. e46939, 2023.
[275] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, “On the evaluations of ChatGPT and emotion-enhanced prompting for mental health analysis,” *arXiv preprint arXiv:2304.03347*, 2023.
[276] P. Lee, S. Bubeck, and J. Petro, “Benefits, limits, and risks of GPT-4 as an ai chatbot for medicine,” *New Engl. J. Med.*, vol. 388, no. 13, pp. 1233–1239, 2023. doi: 10.1056/NEJMsr2214184.
[277] W. H. Organization *et al.*, *Tracking Universal Health Coverage: First Global Monitoring Report*. Geneva, Switzerland: World Health Organization. 2015.
[278] K. Singhal *et al.*, “Large language models encode clinical knowledge,” *arXiv preprint arXiv:2212.13138*, 2022.
[279] S. Yang *et al.*, “Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multiturn dialogue,” *arXiv preprint arXiv:2308.03549*, 2023.
[280] K. Singhal *et al.*, “Towards expertlevel medical question answering with large language models,” *arXiv preprint arXiv:2305.09617*, 2023.
[281] N. H. Shah, D. Entwistle, and M. A. Pfeffer, “Creation and adoption of large language models in medicine,” *JAMA*, vol. 330, no. 9, pp. 866–869, 2023. doi: 10.1001/jama.2023.14217.
[282] A. Amjad, P. Kordel, and G. Fernandes, “A review on innovation in healthcare sector (telehealth) through artificial intelligence,” *Sustainability*, vol. 15, no. 8, pp. 6655, 2023. doi: 10.3390/su15086655.
[283] R. Tang, X. Han, X. Jiang, and X. Hu, “Does synthetic data generation of LLMs help clinical text mining?,” *arXiv preprint arXiv:2303.04360*, 2023.
[284] J. Clusmann *et al.*, “The future landscape of large language models in medicine,” *Commun. Med.*, vol. 3, no. 1, pp. 141, 2023. doi: 10.1038/s43856-023-00370-1.

[285] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "ChatCAD: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023.

[286] Y. Sun, J. He, S. Lei, L. Cui, and C. T. Lu, "Med-MMHL: A multimodal dataset for detecting human-and LLM-generated misinformation in the medical domain," *arXiv preprint arXiv:2306.08871*, 2023.

[287] Y. Wang, X. Ma, and W. Chen, "Augmenting black-box LLMs with medical textbooks for clinical question answering," *arXiv preprint arXiv:2309.02233*, 2023.

[288] I. L. Alberts *et al.*, "Large language models (LLM) and ChatGPT: What will the impact on nuclear medicine be?," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 50, no. 6, pp. 1549–1552, 2023. doi: 10.1007/s00259-023-06172-w.

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

[290] G. Son, H. Jung, M. Hahm, K. Na, and S. Jin, "Beyond classification: Financial reasoning in state-of-the-art language models," *arXiv preprint arXiv:2305.01505*, 2023.

[291] B. Zhang, H. Yang, T. Zhou, M. Ali Babar, and X. Y. Liu, "Enhancing financial sentiment analysis via retrieval augmented large language models," in *Proc. Fourth ACM Int. Conf. AI in Finance*, Brooklyn, NY, USA, 2023, pp. 349–356.

[292] H. Yang, X. Y. Liu, and C. D. Wang, "FinGPT: Open-source financial large language models," *arXiv preprint arXiv:2306.06031*, 2023.

[293] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu and Y. Lu, "Temporal data meets LLM-explainable financial time series forecasting," *arXiv preprint arXiv:2306.11025*, 2023.

[294] Q. Xie, W. Han, Y. Lai, M. Peng, and J. Huang, "The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges," *arXiv preprint arXiv:2304.05351*, 2023.

[295] A. Lykov and D. Tsetserukou, "LLM-BRAIn: AI-driven fast generation of robot behaviour tree based on large language model," *arXiv preprint arXiv:2305.19352*, 2023.

[296] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao, "Large language models for human-robot interaction: A review," *Biomimetic Intell. Robot.*, vol. 3, no. 4, pp. 100131, 2023. doi: 10.1016/j.birob.2023.100131.

[297] Y. Ding, X. Zhang, C. Paxton, and S. Zhang, "Task and motion planning with large language models for object rearrangement," *arXiv preprint arXiv:2303.06247*, 2023.

[298] B. Zhang and H. Soh, "Large language models as zero-shot human models for human-robot interaction," *arXiv preprint arXiv:2303.03548*, 2023.

[299] I. Singh *et al.*, "ProgPrompt: Generating situated robot task plans using large language models," in *2023 IEEE Int. Conf. Robot. Automat. (ICRA)*, London, UK, IEEE, 2023, pp. 11523–11530.

[300] J. Wu *et al.*, "TidyBot: Personalized robot assistance with large language models," *arXiv preprint arXiv:2305.05658*, 2023.

[301] A. Tagliabue, K. Kondo, T. Zhao, M. Peterson, C. T. Tewari, and J. P. How, "REAL: Resilience and adaptation using large language models on autonomous aerial robots," *arXiv preprint arXiv:2311.01403*, 2023.

[302] X. Xiang, Q. Tan, H. Zhou, D. Tang, and J. Lai, "Multimodal fusion of voice and gesture data for UAV control," *Drones*, vol. 6, no. 8, pp. 201, 2022. doi: 10.3390/drones6080201.

[303] Y. Ye, H. You, and J. Du, "Improved trust in human-robot collaboration with ChatGPT," *IEEE Access*, vol. 11, pp. 55748–55754, 2023. doi: 10.1109/ACCESS.2023.3282111.

[304] Y. Zhen *et al.*, "Robot task planning based on large language model representing knowledge with directed graph structures," *arXiv preprint arXiv:2306.05171*, 2023.

[305] 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.

[306] N. Wake, D. Saito, K. Sasabuchi, H. Koike, and K. Ikeuchi, "Text-driven object affordance for guiding grasp-type recognition in multimodal robot teaching," *Mach. Vision Appl.*, vol. 34, no. 4, pp. 58, 2023. doi: 10.1007/s00138-023-01408-z.

[307] C. R. Garrett *et al.*, "Integrated task and motion planning, annual review of control," *Robot. Auton. Syst.*, vol. 4, no. 1, pp. 265–293, 2021. doi: 10.1146/annurev-control-091420-084139.

[308] J. Irons, C. Mason, P. Cooper, S. Sidra, A. Reeson and C. Paris, "Exploring the impacts of ChatGPT on future scientific work," *SocArXiv*, 2023. doi: 10.31235/osf.io/j2u9x.

[309] P. G. Schmidt and A. J. Meir, "Using generative AI for literature searches and scholarly writing: Is the integrity of the scientific discourse in jeopardy?" *Notice Am. Math. Soc.*, vol. 71, no. 1, 2023. pp. 93–103.

[310] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019.

[311] K. Malinka, M. Peresíni, A. Firc, O. Hujnák, and F. Janus, "On the educational impact of ChatGPT: Is artificial intelligence ready to obtain a university degree?" in *Proc. 2023 Conf. Innovat. Technol. Comput. Sci. Educ.*, Turku, Finland, 2023, vol. 1, pp. 47–53.

[312] T. Susnjak, "ChatGPT: The end of online exam integrity?" *arXiv preprint arXiv: 2212. 09292*, 2022.

[313] E. Kasneci *et al.*, "ChatGPT for good? on opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, no. 1, pp. 102274, 2023. doi: 10.1016/j.lindif.2023.102274.

[314] N. L. Rane, "Enhancing the quality of teaching and learning through ChatGPT and similar large language models: Challenges, future prospects, and ethical considerations in education," *TESOL Technol. Studies*, vol. 5, no. 1, pp. 1–6, 2024. doi: 10.48185/tts.v5i1.1000.

[315] S. Küchemann *et al.*, "Are large multimodal foundation models all we need? On opportunities and challenges of these models in education," *EdArXiv*, 2024. doi: 10.35542/osf.io/n7dvf.

[316] D. Tavangarian *et al.*, "Is e-learning the solution for individual learning?" *Electron. J. E-Learn.*, vol. 2, no. 2, pp. 265–272, 2004.

[317] S. Munasinghe *et al.*, "PG-Video-LLaVA: Pixel grounding large video-language models," *arXiv preprint arXiv:2311.13435*, 2023.

[318] C. Zhang, K. Xia, H. Feng, Y. Yang, and X. Du, "Tree species classification using deep learning and RGB optical images obtained by an unmanned aerial vehicle," *J. For. Res.*, vol. 32, no. 5, pp. 1879–1888, 2021. doi: 10.1007/s11676-020-01245-0.

[319] H. Huang, D. Wu, L. Fang, and X. Zheng, "Comparison of multiple machine learning models for estimating the forest growing stock in large-scale forests using multi-source data," *Forests*, vol. 13, no. 9, pp. 1471, 2022. doi: 10.3390/f13091471.

[320] Y. Li, L. Guo, J. Wang, Y. Wang, D. Xu and J. Wen, "An improved sap flow prediction model based on CNN-GRU-BiLSTM and factor analysis of historical environmental variables," *Forests*, vol. 14, no. 7, pp. 1310, 2023. doi: 10.3390/f14071310.

[321] G. Wang, B. Shi, X. Yi, P. Wu, L. Kong, and L. Mo, "DiffusionFR: Species recognition of fish in blurry scenarios via diffusion and attention," *Animals*, vol. 14, no. 3, pp. 499, 2024.

[322] Z. Du, S. Wu, Q. Wen, X. Zheng, S. Lin and D. Wu, "Pine wilt disease detection algorithm based on improved YOLOv5," *Front. Plant Sci.*, vol. 15, pp. 1302361, 2024. doi: 10.3389/fpls.2024.1302361.

[323] C. Zhang *et al.*, "A deep transfer learning toponym extraction and geospatial clustering framework for investigating scenic spots as cognitive regions," *ISPRS Int. J. Geo Inf.*, vol. 12, no. 5, pp. 196, 2023. doi: 10.3390/ijgi12050196.

[324] W. Zhang *et al.*, "ChineseCTRE: A model for geographical named entity recognition and correction based on deep neural networks and the BERT model," *ISPRS Int. J. Geo Inf.*, vol. 12, no. 10, pp. 394, 2023. doi: 10.3390/ijgi12100394.

[325] L. Xu, J. Zhang, C. Zhang, X. Zheng, Z. Du and X. Xue, "Beyond extraction accuracy: Addressing the quality of geographical named entity through advanced recognition and correction models using a modified BERT framework," *Geo-Spatial Inf. Sci.*, pp. 1–19, 2024. doi: 10.1080/10095020.2024.2354229.