Progress in Artificial Intelligence (2025) 14:117–164
https://doi.org/10.1007/s13748-024-00359-4

REVIEW

The role of large language models in agriculture: harvesting the future with LLM intelligence

Tawseef Ayoub Shaikh[1] · Tabasum Rasool[2] · K. Veningston[1] · Syed Mufassir Yaseen[3]

Received: 21 March 2024 / Accepted: 1 December 2024 / Published online: 18 December 2024

Abstract

Significant accomplishments in many agricultural applications during the past decade attest to the fast progress and use of deep learning and machine learning methods in agricultural systems. However, these conventional models have a few drawbacks: They are not generalizable since they are trained on large, costly labeled datasets, require expert expertise to create and maintain, and are often built for specific applications. Significant accomplishments in language, vision, and decision-making tasks across several domains have been shown recently by massive pre-trained models, also known as large models (LMs). Recent years have seen large language models (LLMs) demonstrate remarkable competence in a variety of fields, including natural language processing (NLP), by encompassing different advancements in terms of architecture, training methods, context duration, fine-tuning, multi-modality, datasets, efficiency, benchmarking, and many other. The massive amounts of data used to train these models span many domains and modalities. After training, they can handle a wide range of tasks with less tweaking and less task-specific labeled data. Despite its effectiveness and promising future, agricultural artificial intelligence (AAI) has received less attention than other applications of LLMs. To better understand the problem area and open up new research pathways in this sector, this work aims to examine the possibilities of LLMs in smart agriculture by offering conceptual tools and a technical base. Herein, we delve into the potential applications of large models in agriculture, primarily categorizing them into four categories: Agricultural applications of large language models (LLMs), large vision models (LVMs) for precise agricultural applications, multimodal large language models (MLLMs) and model assessment, and intelligent and precise agriculture using reinforcement learning large models (RLLMs). Further, we review some of the most prominent LLMs, including three famous LLM families (GPT, LLaMA, PaLM), and discuss their characteristics, contributions, and limitations. Next, we evaluate famous LLM evaluation metrics and look at datasets for training, fine-tuning, and evaluation. Finally, we focus our discussion on issues and possible future research directions of LLMs in the agricultural sector. This review article aims to provide academics and practitioners with a panoramic perspective of the field and a quick reference to help them draw out relevant ideas from the extensive summaries of prior publications to broaden their LLM research.

Keywords Language models (LM) · Multimodal large language models (MLLM) · Large vision models (LVM) · Agricultural text classification · Generative pre-trained transformer (GPT) · ChatGPT · Natural language processing (NLP) · Semantic matching (SM)

✉ Tawseef Ayoub Shaikh
tawseef.shaikh@nitsri.ac.in

Tabasum Rasool
tabasumrasool4@gmail.com

K. Veningston
veningstonk@nitsri.ac.in

Syed Mufassir Yaseen
mufassir.syed@mitwpu.edu.in

[1] Department of Computer Science and Engineering, National Institute of Technology (NIT), Srinagar, Jammu & Kashmir 190006, India

[2] NPDF Fellow, Interdisciplinary Centre for Water Research (ICWaR), Indian Institute of Science, Bengaluru, India

[3] Dr. Vishwannath Karad, MIT World Peace Univerity, Kothrud Pune 411038, India

1 Introduction

The close integration of recent information & communication technology (ICT) breakthroughs has revolutionized modern agricultural operations. This has evolved into a thriving area called smart farming, which promises to boost agricultural productivity, efficiency, and product quality. These multi-disciplinary technologies use unmanned aerial/ground vehicles (UAVs/UGVs), image processing, machine learning, big data, cloud computing, and wireless sensor networks (WSNs) [1] to help farmers make informed planting, tending, and harvesting decisions to maximize productivity and profits. However, extracting useful information from varied data sources, particularly imaging data, is difficult. Complex data makes it hard for traditional data mining to find insights. However, deep learning [2] has excelled in processing complicated, high-dimensional data in many applications. DL approaches excel in feature extraction, pattern recognition, and image representation, showing promise in various fields, including agriculture. Among them are weed control, plant disease diagnosis, postharvest quality evaluation, and robotic fruit harvesting [3].

Despite advancements, supervised training, which is fundamental to these methods, requires massive, task-specific, high-quality labeled datasets. Unfortunately, a significant barrier exists for applications with minimal resources due to the enormous time, energy, and money required to acquire and annotate such datasets [4]. The necessity for accurate pixel-level annotations, limitations imposed by biological materials, and imaging settings make this problem more acute in some applications, including weed identification, plant disease diagnosis, and fruit defect detection. Furthermore, there is usually a need to repeat the data gathering and model creation procedure since the gathered datasets aren't very generalizable, even to comparable agricultural areas. This iterative procedure increases the total cost in terms of both time and money and limits the efficiency, scalability, and generalizability of DL frameworks when applied to agricultural applications. Several methods have been suggested to address these problems, like transfer learning [5], few-shot learning [6], and label-efficient learning [7], among others. In particular, transfer learning has been a popular method in many fields, including agriculture, by using DL models that have already been trained on massive image datasets like ImageNet, Microsoft COCO, and PlantCLEF2022 [8]. Then, task-specific adjustments are made to the pre-trained models. In contrast, few-shot learning uses past knowledge (meta-learning techniques) to rapidly train models to adapt to new tasks using a few labeled samples. On the other hand, labeled-efficient learning uses strategies that involve little supervision or none at all to reduce the impact of tedious and time-consuming labeling tasks. However, these approaches' limited applicability to many domains and applications is typically due to their pre-training on data from only one modality. Using multiple data sources to improve its generalizability across many domains and applications is crucial.

Large pre-trained models, often called foundation models (FMs) [9], are artificial intelligence models that can handle many downstream tasks like decision-making, computer vision, speech recognition, and natural language understanding. They train extensively on many different datasets and can produce various outputs. FMs can handle multiple applications from many domains with little fine-tuning and no or little task-specific labeled data. They are often trained on large-scale datasets from varied domains and modalities using self-supervised learning. By using brief text explanations (prompts) that comprise only a few instances, the generative pre-trained transformer language model (GPT-3) [10] can do unexpected tasks without explicit training. Following this, OpenAI created and promoted ChatGPT (OpenAI, 2022), an offshoot of GPT-3, for use in chat-based interactions; this has resulted in a revolutionary shift in NLP, paving the way for more immersive and interactive conversational experiences than ever before. In computer vision, FMs, like the segment anything model (SAM) [11], may handle various downstream segmentation issues using novel image distributions and tasks with zero-shot generalization. They were trained on over 1 billion masks and 11 million licensed and privacy-respecting images. Deepmind Ada [12] introduced FMs to reinforcement learning (RL) that can generalize to new tasks in a few-shot context using a distillation-based teacher-student strategy and a tailored transformer architecture. The use of FMs in artificial intelligence in agriculture has gotten very little attention despite the abovementioned advancements.

1.1 The challenges in the agricultural domain

The importance of agriculture in the global economy is progressively rising, accompanied by an increased awareness of its sustainability. Ahirwar et al. [13] assert that worldwide agricultural food production must be augmented by at least 70% to satisfy the demands of the growing world population. Numerous factors in agriculture impede the consistent augmentation of grain production, including (1) Crop diseases induced by pathogens such as bacteria, fungi, and viruses; (2) The use of unscreened, low-quality seeds resulting in suboptimal crop growth, diminished yield, and increased vulnerability to diseases; (3) Inefficiencies in various agricultural operations, including weeding, planting, irrigation, and harvesting. Agricultural production is experiencing significant economic and output losses. Conventional methods for detecting crop diseases, such as polymerase chain reactions targeting specific deoxyribonucleic acid sequences of pathogens, enzyme-linked immunosorbent assays based

on pathogen proteins, and hyperspectral imaging, are limited by their operational complexity and the necessity for cumbersome equipment. Quality assurance programs use numerous methods to verify seed quality parameters, such as germination and vigor tests, to select high-quality seeds. However, these approaches exhibit constraints regarding temporal overhead, subjectivity, and the destructive assessment of seed quality.

The use of herbicides for weed management in agriculture may have severe environmental consequences, and phytotoxicity responses may result in lower crop quality and yields [14]. Conventional solutions to these activities are wasteful since manned implements are sluggish. As a result, it is vital to establish a quick, simple, and straightforward technique for agricultural workers to meet the highlighted issues. On the other hand, encouraged by rising health awareness, the public has long been concerned about the safety and quality of food, which is associated with agricultural goods. Reduced food losses and improved food safety depend heavily on ongoing crop quality monitoring, particularly disease inspection throughout the crop development. Human civilization has been established across history in significant part by agriculture. Among the many difficulties that have pushed a paradigm transition in contemporary agriculture toward a technologically driven approach in the twenty-first century are limited arable land, water scarcity, climate change, soil degradation, pest and disease outbreaks, and labor shortages [15]. One of the main drivers in solving these challenges has been incorporating new technology into agriculture. New ideas, including "smart agriculture," "precision agriculture," "digital agriculture," "decision agriculture," and "agriculture 4.0," have developed as the agricultural industry explores the potential of digital technology in farming, crop management, and related operations. New agricultural techniques must emerge, especially in less industrialized, less inhabited, and less educated regions. Intelligent facilities and technology may assist in promoting a more sustainable, efficient, and productive agricultural system by easing the load on farmers, raising product quality and market competitiveness, and granting access to agricultural knowledge and information.

Artificial intelligence and the internet of things (IoT) are among the most often discussed and studied modern technologies for prospective use in agriculture. By combining digital data from the physical world with digital data via the IoT and high-tech sensors, advanced technologies let one create digital representations of real-world items and environments [16]. In the agricultural field, it offers game-changing power and opens the path for risk analysis, resource optimization, decision support, equipment monitoring, and precision farming. Two artificial intelligence applications, computer vision and prediction systems, have also tremendously improved farming. Creative applications have emerged in various sectors thanks to better artificial intelligence algorithms.

1.2 Large language-based artificial intelligent solutions

Large models are often used in agriculture as an effective analytical tool. Large models have performed very well in agricultural data analysis, pest and disease control, precision agriculture, and other applications. However, it still confronts several challenges, including difficulties getting agricultural data, poor model training efficiency, distribution shift, and plant blindness [17]. Big models such as DALL, GPT, and the diffusion model do quite well regarding visual material, multi-modal analysis, and natural language. These versions have been refined to satisfy various industries [15]. Legal documents, such as those related to plant protection measures, may be appropriately described by LLMs, which are also adept at explaining the application of special area regulations. The LLM's use of probabilistic tokenization makes this possible by lowering the size of the datasets. Advanced LLMs like GPT-3 have incorporated data purification and reinforcement learning (RL) with human input. This model produces more understandable results by tweaking specific parameters. The authors et al. [9] brought up German legislation, known as the "Plant Protection Application Regulation," while querying GPT-3.5. This 23-page regulation implementing the plant protection act has complex and cross-referenced interrelationships. GPT-3.5 uses bullet points to deliver concise, straightforward replies to particular rule inquiries; the underlying interrelationships have already been unlocked. Regrettably, laws are often modified, and a renewed commitment to understanding the LLM is needed to deliver an accurate and valid answer. A high-quality model demands a significant amount of effort and money. However, LLMs have significant applications in farming. Starting with (a) Consulting and assistance, (b) Automated documentation, (c) Explanation and education, (d) Interpretation of ML results and forecasts, (e) Personalized advisory to farmers, (f) Eligibility for schemes, (g) Monitoring the problems faced by farmers in real-time, and proactively managing those at state level (h) Personalized training in agriculture—educational content in agriculture (i) Optimizing crop rotation (j) Market intelligence for better pricing, etc. Figure 1 below presents a wide range of applications large language models can offer to the agriculture sector.

The notion of agricultural advise is incorporated in consulting and assistance. This may help identify the best time, activity, and tool to optimize yield in the field. Data gathered by farm management systems and machine trackers is often machine-readable; automated documentation describes how this data is converted to human-readable language. Instructional resources, such as guides, tutorials,

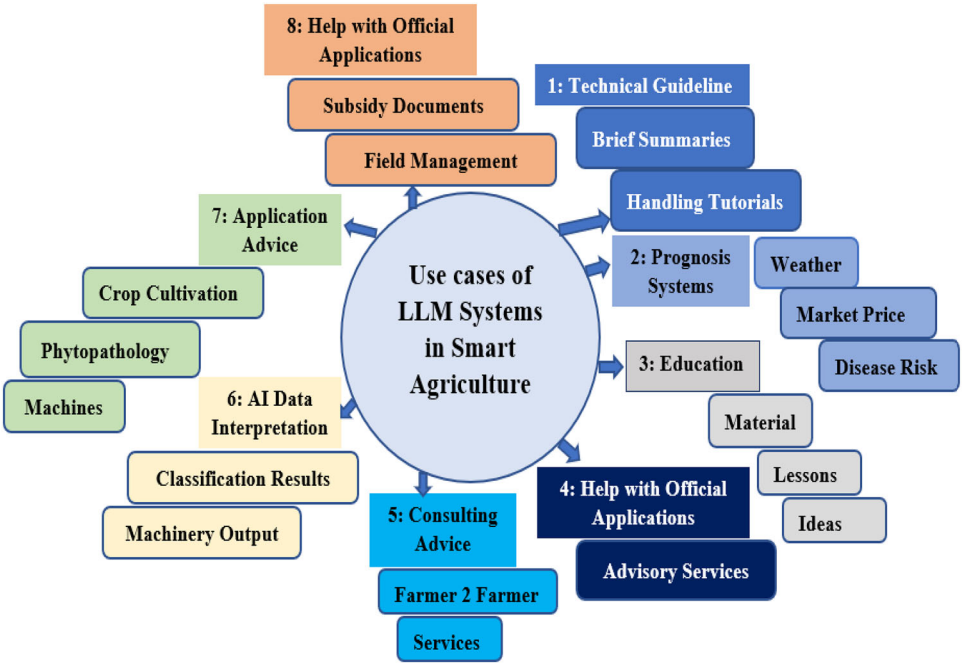

Fig. 1 Applications of large language models in agriculture, such as automated reporting, technical guidelines, application decision support, textual processing, official application assistance, and consulting

films, and books, are intended to be created automatically for self-study. Beginning with a less technical and more understandable introduction is advisable when dealing with digital technology. Finally, machine learning results and forecasts enable context-based interpretation or direct decision support [18]. Consultations with the system provide farmers contextual assistance, while individualized advisory services offer excellent recommendations. Farmers face significant challenges when they do not have someone to speak to at every stage of crop management. Subsidy distribution has progressively replaced the advice services formerly given by government agriculture officials. Furthermore, the generative AI may learn from a farmer's questions and responses, enabling it to build and improve itself in real-time. Some Indian firms with the appropriate infrastructure have moved quickly to start three pilot initiatives. In Odisha, Samagra launched the pilot Ama Krush. In addition to DigitalGreen and Gooey.ai, Apurva.ai has launched Farmer-chat and a pilot program for National Digital Extension [18]. The development of commodities that are disease-resistant, high-yielding, and adaptable to climate change is possible with the aid of artificial intelligent-based technologies. The diverse applications of artificial intelligence-based solutions to the agriculture sector are portrayed in Fig. 2.

People eligible for schemes may find learning more about government websites helpful. There is good support for generative artificial intelligence trained with government programs' data. After that, it might help customers find the right government program. A critical part of smart farming methods is keeping an eye on the problems farmers are having in real-time and taking steps to solve them [19]. Because

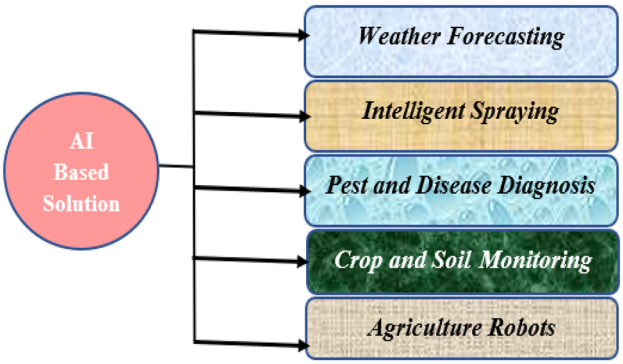

Fig. 2 Artificial intelligence-based solutions in smart and precision agriculture

farmers have so many issues, most states give them a phone line they can call for help. Voice-based machine learning and call summary make it possible to automatically track and sort the daily call numbers into the different types of problems farms have. At this point, generative AI provides a teaching agent that answers each farmer's unique and personalized question. A special LLM for agriculture could use data from all the data sources and data from different institutions in the field. After that, this virtual agent can be improved to be a personalized teacher for farmers. Two essential things crop rotation does are keep the soil healthy and prevent diseases. LLMs can tell the best ways to rotate crops by looking at data from the past. Each farmer needs to understand how market prices and trends work. LLM studies of market data can give instant information about price trends, changes in demand, and customer tastes [20].

We aim to offer a comprehensive analysis of large models, starting with a systematic summary of the history of large models (large language models, large vision models, multimodal large language models, and reinforcement large language models), large models in other fields, the importance of large model for agriculture. Subsequently, we introduce many applications of large models in agriculture. As such, this study aims to explore the potential of developing and applying LLMs for agricultural applications. More specifically, we first review recent large language models in the general computer science (CS) domain and categorize them into four categories: large language models (LLM) in agricultural applications, large vision models (LVM) for farm applications, multimodal large language models (MLLM) and model assessment, and reinforcement learning foundation models (RLFMs) in agriculture. Subsequently, we outline the process of developing large agriculture foundation models (ALFMs) and discuss their potential applications in smart agriculture. Moreover, because large models are a relatively new technological means, we outline some solutions based on their ethical and responsibility aspects. Finally, we summarize the current challenges and future directions of large models and conclude the effectiveness of their implementation in the agricultural domain. This comprehensive examination of agriculture's large models can serve as a valuable resource for newcomers to the field and experienced researchers seeking to innovate within the agricultural space. The article details the difficulties and dangers of creating LLMs, including model training, validation, and deployment. Agricultural LLM scan newbies and seasoned researchers alike will find this in-depth analysis invaluable for agricultural innovation. To the best of our knowledge, this is the first comprehensive review paper on large language models in the smart agriculture domain. The novel ingredients of this contribution are given below.

1.3 Contributions

To this end, this paper presents a brief history and analysis of architectures, applications, evaluations, and security issues in large language models. The contributions of this paper are:

(a) We provide a detailed background of different types of LLMs and their general architecture.
(b) A comprehensive literature survey about LLMs related to various computer science fields. A state-of-the-art review, analysis, and comparison of security issues for LLMs.
(c) Motivated by the progress of large pre-trained language models like ChatGPT, we conducted a preliminary study on agricultural text classification.
(d) The applications of LLMs in smart and precision agriculture are discussed. More specifically, the applications are categorized into four categories: large language models (LLM) in agricultural applications, large vision models (LVM) in agricultural applications, multimodal large language models (MLLM) and model assessment, and reinforcement learning foundation models (RLFMs) in agriculture. These four sub-categorizations are further divided into various sub-sections.
(e) We have also provided detailed application scenarios of LLMs in other fields like medicine, education, science, mathematics, law, finance, and coding.
(f) An analysis of LLMs security requirements and challenges, possible solutions, and areas for future research are discussed.

1.4 Layout of the paper

The remaining paper is divided into six additional sections, as illustrated in Fig. 3 below. Section 2 begins with a detailed understanding of the history, applications, and current status of LLM models. A detailed overview of the various diverse large language-based models, large language model families, their architectures, taxonomy, and qualitative analysis is carried out in this Section. Section 3 discusses LLMs' applications in smart and precision agriculture, medicine, education, science, mathematics, law, finance, and coding. More specifically, in agriculture, we further categorized the applications into four more domains, ranging from large language models in agricultural applications, large vision models in agricultural applications, multimodal large language models and model assessment, and reinforcement learning foundation models (RLFMs) in agriculture. Various benchmark datasets for evaluating LLM models with their detailed discussions are at the core of Sect. 4. Section 5 presents multiple challenges and dark sides of LLMs related to security deep fakes, and a discussion of possible solutions is also provided. Finally, the paper ends with new research findings listed as Conclusions in Sect. 6.

2 Understanding language-based systems: a historical perspective

This section gives an overview of the history of large language-based systems to help understand their complexities, needs, architectures, families, types, and applications. The artificial intelligence field has grown and slowed down many times since it began after World War II. During these stages, there are often changes in how universities and businesses pay for artificial intelligence research.

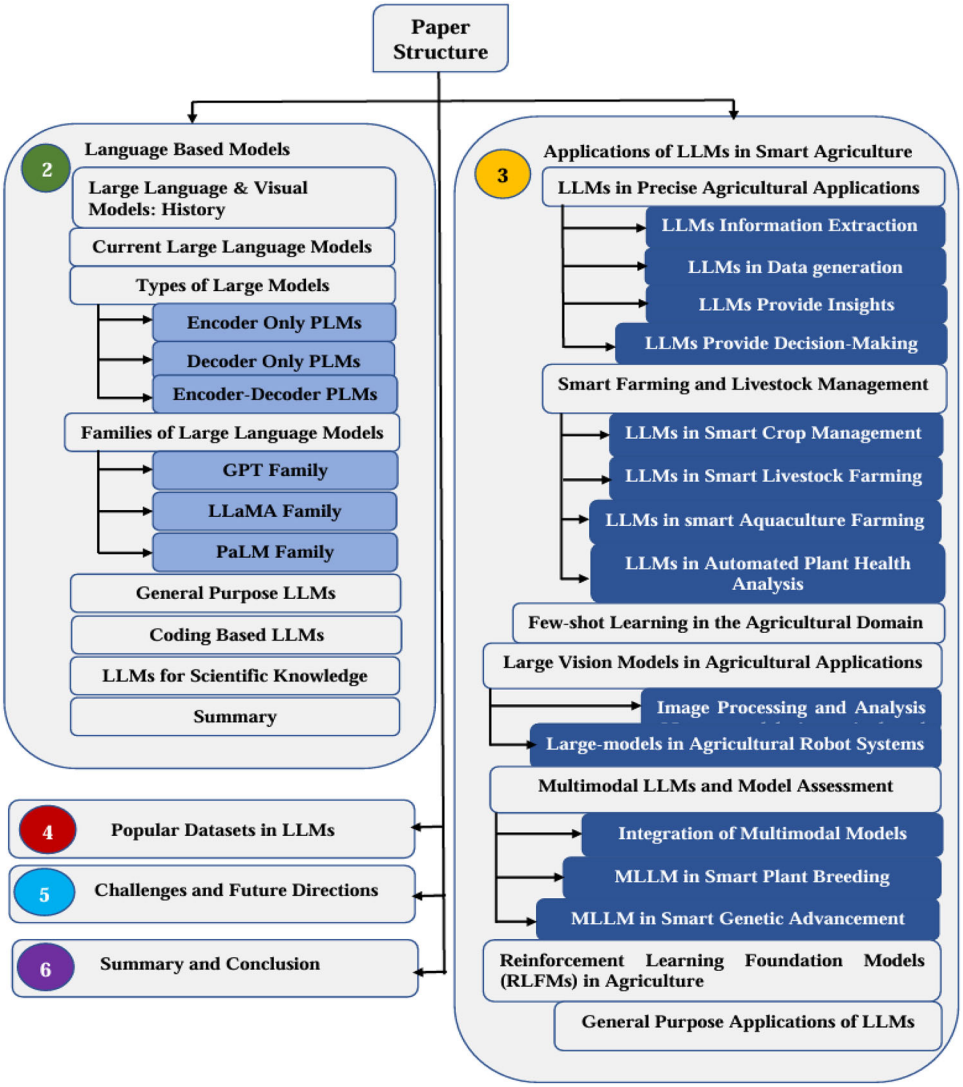

Fig. 3 Complete structure of the paper

2.1 The history of large language and large visual models

The main goal of artificial intelligence is to make machines that can learn and think like humans, similar to how humans can see and speak. For now, natural language processing (NLP) and computer vision (CV) are also at the center of study on big models. NLP is where LLM and large visual models (LVM) get their ideas, and we can break their growth into four stages:

Statistical language models (SLM): Statistical language models learn the probability distribution of words using n-gram and other classic statistical methods and specific linguistic rules during training. In order to provide a solution that can compete effectively with the established n-gram language models, it is widely considered that the quantity of data and the capacity of a specific estimate method to handle a high amount of training are crucial factors [21]. One example of the widespread usage of SLMs in natural language processing is the static analysis that Raychev et al. [21] developed for fixing faulty code, which is both easy and scalable. However, there are three problems with n-gram models. First, the amount of memory required to compute and count parameters grows directly proportional to the value of n. The size of n may be constrained using the Markov assumption. Second, models using n-grams cannot exchange semantically related vocabulary or prefix data. One way to convert text to a vector format is by using word embedding, which can even be problematic in cases of data sparsity. One way to address this issue is using data smoothing, backoff, and interpolation [22]. Also, data sparsity isn't a concern for neural network models.

- Neural language models (NLM): To simulate language, they use a variety of neural networks, and when compared to SLMs, neural language models are more successful

[23]. Feedforward neural networks and recurrent neural networks (RNN) are used in continuous space language modeling to address data sparsity in n-gram models. These neural networks allow the model to learn features and continuous representations automatically. Bengio et al. [24] proposed the first feedforward neural network language model (FFNNLM). This model can solve the dimensionality problem by learning a word's distributed representations. Long short-term memory recurrent neural networks were applied to language models by Sundermeyer et al. [25] by proposing LSTM-RNNLM. The issue of long-term reliance on language model learning was resolved by adding three gate structures to the LSTM memory unit. These gate structures included input, output, and forget gates. The purpose of these gate structures was to manage the flow of information.

- Pre-trained language models (PLM): Two different paradigms may be used to classify pre-trained language models: feature-based and fine-tuning. The feature-based approach approaches pre-training as a feature extraction process, trains model parameters on large-scale corpora, and encodes them as fixed features for use by downstream models in collective tasks. Examples such as ElMo, a pre-training bidirectional LSTM (BiLSTM) proposed by Peters et al. [26], are typical examples. Since LSTM models sentences, it can only take into account the contextual information that comes before the present sentence and cannot take into account the contextual information that comes after it. Additionally, BiLSTM uses reverse networks, which can simultaneously consider contextual details before and after, resulting in improved sequential data processing. The fine-tuning paradigm, which is the current dominant paradigm and has higher flexibility than the feature-based paradigm, transfers the parameters of the whole model to actions performed farther down the line. The bidirectional encoder representations from transformers (BERT) and the generative pre-trained transformer (GPT) are the models that illustrate the fine-tuning process. In 2017, the research team at Google published a model called Transformer [27], which includes a self-attention mechanism. At the same time, OpenAI developed GPT, which was built on the architecture of Transformer [27]; GPT got almost flawless training outcomes by doing preliminary training on large text datasets and then fine-tuning the parameters. The BERT algorithm was developed by pre-training bidirectional language models with specifically tailored pre-training tasks on large unlabeled corpora. The effectiveness of these pre-trained context-aware word representations as general-purpose semantic features is high, resulting in a considerable improvement in the performance of natural language processing tasks. Additionally, due to the tremendous acceleration of model training that Transformer provides, it has increasingly become the essential architecture for LLMs.
- Large language models (LLM): LLM is a language model with billions or more parameters. Large models have capabilities that small models lack, referred to as the emergent capacities of LLMs. This is a prominent differentiating characteristic between LLMs and PLMs. OpenAI researchers found that bigger models would consistently demonstrate superior performance and significantly enhanced sampling efficiency compared to previous iterations [28]. Numerous contemporary research has trained large-scale PLMs and discovered that, in contrast to smaller PLMs, large-scale PLMs have distinct behaviors and remarkable capabilities in addressing various complicated problems [28]. This refers to the emergent capabilities of LLMs, as previously noted. For example, GPT-3's contextual learning capability may provide anticipated outputs for test samples by completing word sequences from the input text without requiring further training or gradient adjustments, a feat unattainable by GPT-2. Consequently, the research community designates these large-scale PLMs with enhanced functionalities as LLMs [29].

Large vision models (LVM), on the other hand, are models linked to computer vision (CV). The investigation into vision models first concentrated on superficial image feature extraction algorithms, such as scale-invariant feature transform, histogram of directed gradients, and other techniques, but encountered considerable restrictions. In 2012, AlexNet [30] attained a significant breakthrough in the ImageNet Large Scale Visual Recognition Challenge, catalyzing a proliferation of convolutional neural networks (CNN) for visual modeling [31–33]. The advancement of deep learning has led to the sequential introduction of deep residual networks such as VGGNet, GoogLeNet, and ResNet, which enhanced the efficacy of image classification, object recognition, and semantic segmentation. The proliferation of the internet facilitated the use of extensive image collections for training vision models. Faster R-CNN [34], YOLO [35], and Mask R-CNN [36] were developed sequentially. In recent years, transformers have been used in the field of LVM, with the emergence of vision transformers (ViT) [37] and DALL-E [38]. These models use a self-attention mechanism with a generative adversarial network to exhibit robust proficiency in image classification and creation tasks.

Besides the LLM mentioned above and LVM, multimodal large language models (MLLM) are also a focal point of study in large language models. Large language models excel at text-based activities; nevertheless, they are challenging to comprehend and manage with other data kinds. Large vision models excel in computer vision, yet there is a lack of information on the analysis outcomes, which imposes some

constraints on users. MLLMs [39] amalgamate several data modalities, including images, text, language, and audio. It encompasses the benefits of LLMs and LVMs and mitigates their limits by merging several modalities, facilitating a more holistic comprehension of diverse material. The advancements in MLLMs have established new pathways for AI, enabling binary computers to comprehend and subsequently analyze diverse data formats.

2.2 The currently developed large models

Several industry experts have discovered that large models may yield significant sector advancements. Several businesses have begun sequentially allocating human resources, materials, and financial investments to develop a substantial model appropriate for industrial applications capable of executing certain professional activities. Table 1 illustrates that the predominant large models are mostly LLMs and MLLMs, whereas LVMs constitute a minority. Numerous large language models are engineered to create chatbots BLOOM [41], PaLM2 [42], ERNIE Bot or to execute other natural language processing (NLP) tasks, such as text classification, machine translation, and sentiment analysis OPT [43]. Certain researchers are dissatisfied with NLP tasks; hence, they have included visual capabilities to allow the model to respond to inquiries based on images like Minipt-4 [44]. This kind of model is called a large vision-language model (LVLM). While LVLM fulfills certain functions and significantly advances large models toward artificial general intelligence (AGI), it remains insufficient to enable machines to replicate human cognition and perform a broad spectrum of general tasks via transfer learning and various modalities without attaining the model's multimodality [45]. Certain extensive models have included multimodality, allowing them to evaluate multiple forms of information like GPT-4 [46], LLaMA [47], Gemini [48], ImageBind [49] and engage with users.

However, since many existing models are generic and their training datasets are too diverse, they cannot offer a suitable solution to questions in specific professional disciplines. According to Goertzel, a system does not need infinite generality, adaptability, or flexibility to be classified as AGI [53]. As a result, several researchers refined and altered existing large models before releasing new large models tailored to a particular subject. BloombergGPT has shown exceptional performance on general LLM benchmarks and outperforms equivalent models on financial jobs. Huawei's panguLM meteorological model can forecast gravitational potential, humidity, wind speed, temperature, and pressure within an hour to 7 days. PaLM-E integration with robots may do various tasks, including visual question answering, sequential robotic manipulation planning, and captioning. OCEANGPT is a specialist in multiple marine scientific jobs [54]. It demonstrates greater knowledge competence for ocean research activities and develops preliminary embodied intelligence skills for ocean engineering. PMC-LLaMA is a pioneering open-source medical language model that outperforms ChatGPT and LLaMA-2 on various medical benchmarks using fewer parameters.

2.3 Pathway to large language model families

In this subsection, we review early pre-trained neural language models as they are the base of LLMs. Then, we focus our discussion on three families of LLMs: GPT, LlaMA, and PaLM. Table 1 provides an overview of some of these models and their characteristics.

2.3.1 Early pre-trained neural language models

Language modeling using neural networks was pioneered in the early developmental days of LLM research. Bengio et al. [64] created one of the first neural language models (NLMs) similar to n-gram models. Then, [65] et al. successfully used NLMs for machine translation. Mikolov's introduction of RNNLM (an open-source NLM toolkit) [66] contributed substantially to the popularity of NLMs. Later, NLMs based on recurrent neural networks (RNNs) and their derivatives, such as long short-term memory (LSTM) and gated recurrent unit (GRU) [67], were popular for a variety of natural language applications, including machine translation, text production, and text categorization. The transformer architecture is then invented, marking another milestone in the development of NLMs. Transformers enable much more parallelization than RNNs by using self-attention to compute in parallel for each word in a sentence or document an "attention score" to model the influence each word has on another, allowing for the efficient pre-training of large language models on large amounts of data on GPUs. These pre-trained language models (PLMs) may be fine-tuned for future activities. We categorize early popular transformer-based PLMs based on their neural topologies into three types: encoder-only, decoder-only, and encoder-decoder models.

2.3.1.1 Encoder-only PLMs The encoder-only types sound precisely like they only have an encoder network. These models were first created for jobs that require understanding language, like text classification, where they have to guess what class a piece of text belongs to. Some examples of encoder-only models are BERT and its versions, such as RoBERTa, ALBERT, DeBERTa, XLM, XLNet, and UNILM.

It is one of the most popular encoder-only language models [68]. Its name comes from the fact that it uses bidirectional encoder representations from transformers. BERT is made up of three parts: (1) An embedding module that takes in text and turns it into a series of embedding vectors; (2) A stack of

Table 1 The currently popular large models

llm name	Release date	Open source	Models	Information	References
Chinchilla	March 29th, 2022	No	LLM	Training a large language model with optimal computational utilization	[40]
OPT	May 2rd, 2022	Yes	LLM	The full name of OPT is open pre-trained transformer language models, which means "open pre-trained transformer language models."	[43]
BLOOM	November 9th, 2022	Yes	LLM	A decoder-only model based on Transformer architecture	[41]
Minigpt-4	April 20th, 2023	Yes	LLM	This model can understand images and text and respond to user instructions	[44]
LLaMA Adapter V2	August 28th, 2023	Yes	LLM	A Parameter Efficient Visual Instruction Model	[47]
PMC LLaMA	April 27th, 2023	Yes	LLM	Inject medical knowledge into existing LLM using 4.8 million biomedical academic papers	[50]
PaLM-E	March 6th, 2023	No	LLM	PaLM-E is a large language model with only a decoder	[51]
PaLM2	May 11st, 2023	No	LLM	PaLM2 is a neural network-based language model that is considered one of the most advanced language models currently available	[42]
ERNIE Bot	4.0 October 17th, 2023	No	LLM	ERNIE Bot is a new generation of Baidu's big language model for knowledge enhancement	–
Qwen-7B	August 3rd, 2023 Cloud	Yes	LLM	A super large language model launched by Alibaba	[52]
IFLYTEK SPARK	May 6th, 2023 field	No	LLM	IFLYTEK SPARK is a new generation of cognitive intelligence model with Chinese as its core	–
Bloomberg GPT	March 30th, 2023	No	LLM	A LLM for the financial field	[53]
OCEANG PT	October 3rd, 2023	Yes	LLM	A LLM for ocean science tasks	[54]
InternIma ge	April 17th, 2023	Yes	LVM	A large visual model based on deformable convolution	[55]
PanguCV LM	April 25th, 2021	No	LVM	PanguCVLM is a model that utilizes a large model network to simulate and automate human visual processes	–
LLaVA	April 17th, 2023	Yes	MLLM	LLaVA, a new large multimodal model	[56]
Instruct BLIP	June 15th, 2023	Yes	MLLM	Instruct BLIP model achieves state-of-the-art zero sample performance on various visual language tasks	[57]
Visual ChatGPT	March 8th, 2023	Yes	MLLM	The proposal of Visual ChatGPT opened the door to the connection between ChatGPT and VFM, enabling ChatGPT to handle complex visual tasks	[58]
mPLUG Owl	April 27th, 2023	Yes	MLLM	A multimodal dialogue generation model similar to miniGPT-4 and LLaVA, with image viewing and chat functionality	[59]
VisualGL M-6B	May 17th, 2023	Yes	MLLM	VisualGLM-6B is an open-source multimodal dialogue language model that supports images, Chinese, and English	[60]

Table 1 (continued)

llm name	Release date	Open source	Models	Information	References
ImageBind	May 9th, 2023	Yes	MLLM	ImageBind is the first artificial intelligence model to bind information from six modes	[49]
MultiMod al-GPT	May 8th, 2023	Yes	MLLM	MultiModal GPT can follow various human instructions, such as generating detailed instructions, calculating the number of objects of interest, and answering general questions from users	[61]
GPT-4	March 14th, 2023	No	MLLM	GPT-4 is the latest pre-trained model launched by OpenAI, and it is aimed at improving the performance of machine learning models using NLP technology	[46]
Skywork	April 17th, 2023	No	MLLM	Skywork is a series of large models developed by the Kunlun · Skywork team	[62]
Gemini	December 6th, 2023	No	MLLM	Gemini is an artificial intelligence multimodal large language model launched by Google DeepMind	[48]
Sora	February 15th, 2024	No	MLLM	Sora is an AI model that can create realistic and imaginative scenes from text instructions	[63]

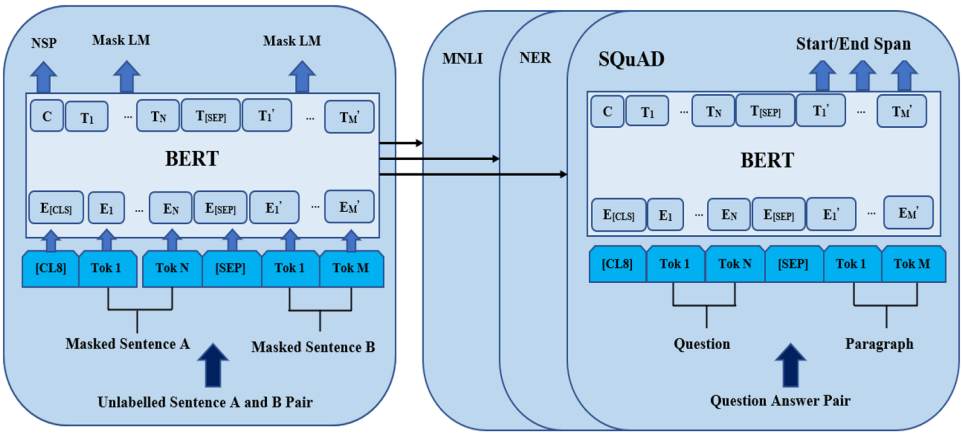

Fig. 4 Overall pre-training and fine-tuning procedures for BERT

transformer encoders that turns the embedding vectors into contextual representation vectors; and (3) A fully connected layer that turns the representation vectors into one-hot vectors at the top level. BERT has already been taught to do masked language modeling (MLM) and predict the following line. A pre-trained BERT model can be improved by adding a classifier layer. This can be done for many language learning tasks, such as text classification, question answering, and language reasoning. Figure 4 shows a broad look at the BERT structure. When BERT came out, it made a big difference in the state of the art for many language learning tasks. This led the AI community to create similar encoder-only language models based on BERT.

Using a series of model design decisions and training techniques, including changing a few essential hyperparameters, eliminating the next-sentence pre-training aim, and training with considerably bigger mini-batches and learning rates, RoBERTa [69] dramatically increases the robustness of BERT. To reduce memory use and speed up BERT training, ALBERT [70] employs two parameter-reduction strategies: (1) Dividing the embedding matrix into two smaller matrices and (2) Using repeating layers divided across groups. DeBERTa (decoding enhanced BERT with disentangled attention) enhances the BERT and RoBERTa models using two innovative methods. The first is the disentangled attention mechanism, in which the attention weights between words are calculated using disentangled matrices on their relative locations and contents, respectively, and each word

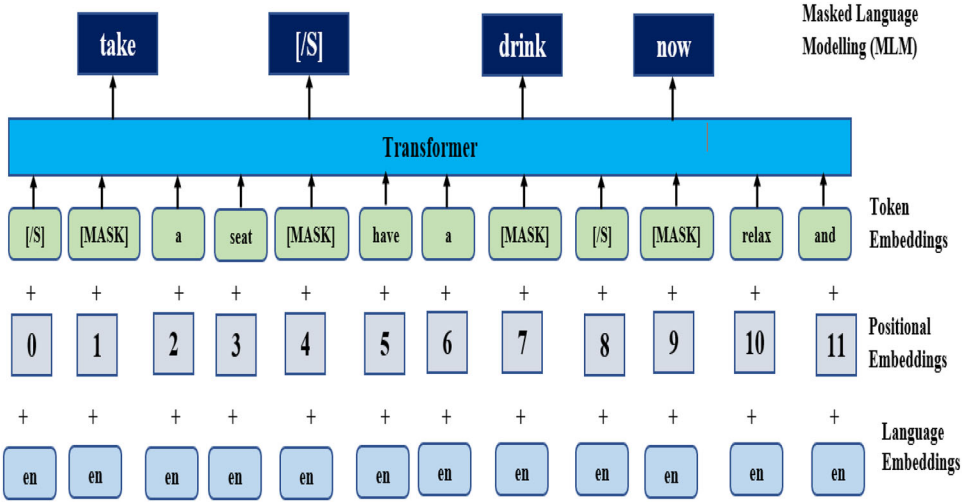

Fig. 5 Cross-lingual language model pretraining. The MLM objective is similar to BERT but with continuous streams of text as opposed to sentence pairs

is represented by two vectors that encode its position and content, respectively. Second, to forecast the masked tokens in model pre-training, absolute locations are included in the decoding layer using an improved mask decoder. Furthermore, a new virtual adversarial training technique is used for fine-tuning to enhance the generalization of models. UELECTRA [71] employs a novel pre-training task known as replaced token detection (RTD), which has been experimentally shown to be more sample-efficient than MLM. Instead of concealing the input, RTD modifies it by replacing specific tokens with plausible replacements drawn from a tiny generator network. Then, rather than developing a model to predict the original identities of the corrupted tokens, a discriminative model is trained to determine whether a produced sample replaced a token in the corrupted input. RTD is more sample-efficient than MLM since it is defined over all input tokens rather than just the tiny fraction masked. XLMs [72] adapted BERT to cross-lingual language models by using two methods: (1) An unsupervised technique that exclusively uses monolingual data, and (2) A supervised method that uses parallel data and a new cross-lingual language model goal, as shown in Fig. 5.

When XLMs were first suggested, they yielded cutting-edge results in cross-lingual categorization and unsupervised and supervised machine translation. Encoder-only language models also use the benefits of auto-regressive (decoder) models during model training and inference. Two examples are XLNet and UNILM. XLNet [73] is built on Transformer-XL, which was pre-trained using a generalized autoregressive technique that allows for bi-directional context learning by maximizing the expected probability over all factorization order permutations. UNILM (UNIfied Pre-trained Language Model) [74] is trained on three language modeling tasks: unidirectional, bidirectional, and sequence-to-sequence prediction. This is accomplished by using a shared Transformer network and unique self-attention masks to regulate the context on which the prediction is conditioned. The pre-trained model may be fine-tuned for natural language comprehension and generating tasks.

2.3.1.2 Decoder-only PLMs OpenAI produced two of the most extensively used decoder-only PLMs: GPT-1 and GPT-2. These models serve as the basis for more powerful LLMs, such as GPT-3 and GPT-4. GPT-1 [75] illustrates for the first time that solid performance across a wide variety of natural language tasks may be attained using Generative Pre-Training (GPT). The transformer model is trained on a varied corpus of unlabeled text using self-supervised learning (i.e., next word/token prediction), followed by discriminative fine-tuning on each downstream job (with considerably fewer samples), as shown in Fig. 6. GPT-1 provides the groundwork for future GPT models, with each iteration refining the architecture and attaining higher performance on diverse language tasks. GPT-2 [76] demonstrates that language models may learn to do specific natural language tasks without explicit supervision when trained on a vast WebText dataset, including millions of websites. The GPT-2 model is based on the GPT-1 model designs, with a few modifications: Layer normalization is moved to each sub-block input, additional layer normalization is added after the final self-attention block, initialization is modified to account for accumulation on the residual path and scaling the weights of residual layers, vocabulary size is increased to 50.25, and context size is increased from 512 to 1024 tokens.

2.3.1.3 Encoder-decoder PLMs Raffle et al. [77] show that almost all NLP tasks can be cast as sequence-to-sequence generation tasks. Thus, an encoder-decoder language model, by design, is a unified model in that it can perform all-natural language understanding and generation tasks. T5 [77] is a Text-to-Text Transfer Transformer (T5) model where

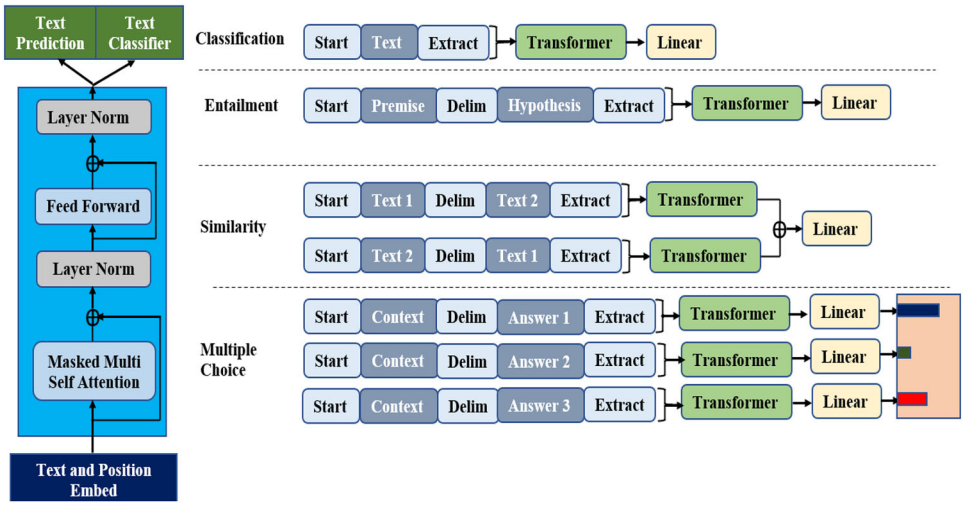

Fig. 6 High-level overview of GPT pretraining and fine-tuning steps. Courtesy of OpenAI

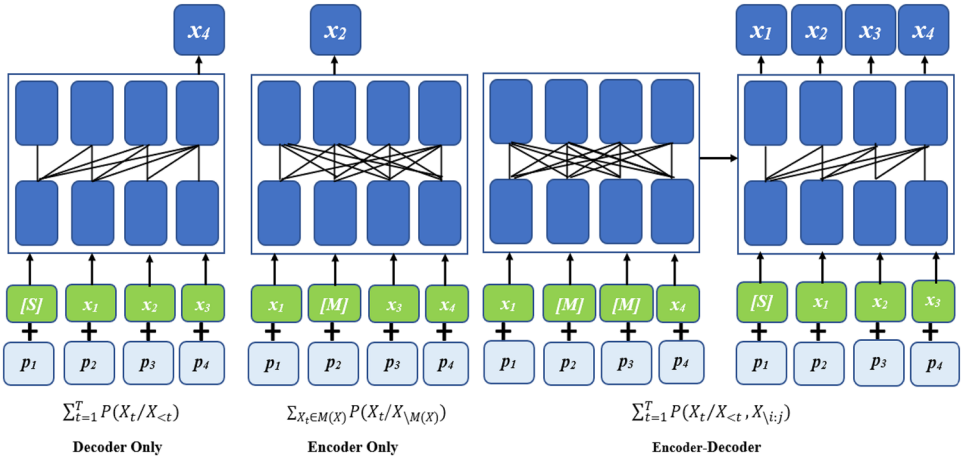

Fig. 7 An illustration of the existing prevalent pre-training frameworks, where x is the original sentence, $x_t(t = 1,2,\ldots,T)$ is the t_{th} token, T is the sequence length, and $M(x)$ is the set of masked tokens in x. S denotes the start token embedding of a sequence. *p1, p2, p3, and p4* denote the position embeddings of the first to fourth tokens. P is the conditional probability. i and j indicate the start and the end indices of input tokens of the encoder, respectively

transfer learning is effectively exploited for NLP by introducing a unified framework in which all NLP tasks are cast as text-to-text generation tasks. mT5 [78] is a multilingual variant of T5, pre-trained on a new Common Crawl-based dataset consisting of texts in 101 languages. MASS (MAsked Sequence to Sequence pre-training) [79] adopts the encoder-decoder framework to reconstruct a sentence fragment given the remaining part of the sentence. The encoder takes a sentence with a randomly masked fragment (several consecutive tokens) as input, and the decoder predicts the masked fragment. In this way, MASS jointly trains the encoder and decoder for language embedding and generation, respectively. BART [80] uses a standard sequence-to-sequence translation model architecture. It is pre-trained by corrupting text with an arbitrary noising function and then learning to reconstruct the original text. Figure 7 summarizes the existing pre-training frameworks, which can be classified into three categories: transformer decoders, transformer encoders, and transformer decoder–encoders.

2.3.2 Large language model families

Large language models (LLMs) are transformer-based PLMs with tens to hundreds of billions of parameters. Compared to the PLMs discussed above, LLMs are substantially bigger in model size and display greater language comprehension and generation and emergent skills absent from smaller-scale models. This section examines LLMs, briefly discussing their structures, training goals, pipelines, datasets, and fine-tuning options. This section also examines other LLMs, briefly discussing their structures, training goals, pipelines, datasets, and fine-tuning options. We have divided these numerous LLM models into seven categories, starting with three LLM families: GPT, LLaMA, and PaLM, as shown in Fig. 8. Then, it is followed by pre-trained general-purpose LLMs, pre-trained coding-based LLMs, pre-trained LLMs for scientific knowledge, and lastly, other representatives of LLMs.

2.3.2.1 The GPT family OpenAI created generative pre-trained transformers (GPT), a series of decoder-only transformer-based language models. This family includes

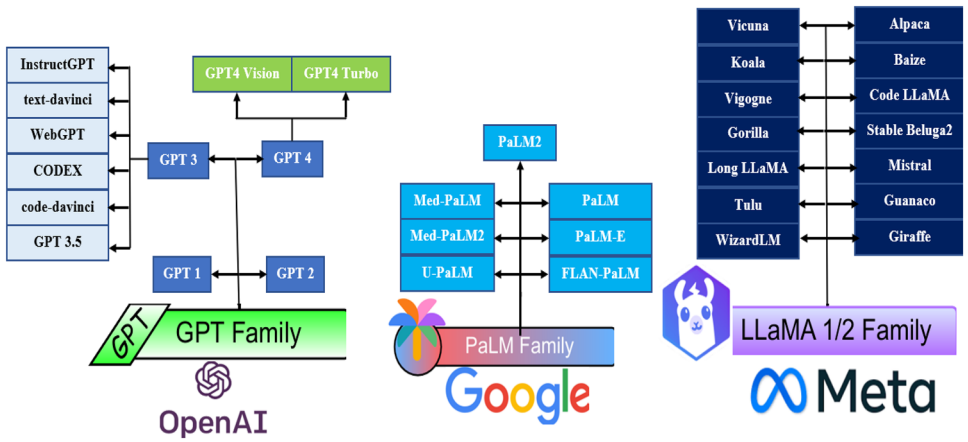

Fig. 8 Popular large language models families

GPT-1, GPT-2, GPT-3, InstrucGPT, ChatGPT, GPT-4, CODEX, and WebGPT. Although early GPT models, such as GPT-1 and GPT-2, were open-source, more contemporary versions, such as GPT-3 and GPT-4, are closed-source and only accessible via APIs. The early PLM segment covered the GPT-1 and GPT-2 models. GPT-3 [81] is a pre-trained autoregressive language model with 175 billion parameters. GPT-3 is commonly regarded as the first LLM since it is much bigger than prior PLMs and displays emerging skills not seen in previous smaller PLMs. GPT 3 demonstrates the emergent capacity of in-context learning, which implies it may be applied to any downstream tasks without requiring gradient updates or fine-tuning, with functions and few-shot demos provided just by text interaction with the model. GPT-3 performed well on several NLP tasks, including translation, question-answering, and cloze tasks, as well as those that required on-the-fly reasoning or domain adaptation, such as unscrambling words, employing a new term in a sentence and 3-digit arithmetic. They also demonstrate higher language comprehension and generation and emergent skills, which are not evident in smaller-scale models.

CODEX [82], launched by OpenAI in March 2023, is a general-purpose programming model capable of parsing natural language and producing code in response. CODEX is a descendant of GPT-3 that has been fine-tuned for programming applications using code corpora from GitHub.

WebGPT [83] is another grandchild of GPT-3, fine-tuned to answer open-ended inquiries using a text-based web browser, allowing users to search and traverse the web. Specifically, WebGPT is trained in three stages. The initial step is for WebGPT to learn to emulate human browsing activities from human demonstration data. Then, a reward function is trained to anticipate human preferences. Finally, WebGPT is improved to maximize the reward function via reinforcement learning and rejection sampling. InstructGPT [84] is suggested to align language models with user intention on various activities by fine-tuning using human input, thereby enabling LLMs to follow anticipated human instructions. A dataset of labeler demonstrations of the required model behavior is gathered, starting with a set of labeler-written prompts and prompts sent over the OpenAI API. GPT-3 is then adjusted on this dataset. After that, a collection of human-ranked model outputs is gathered to apply reinforcement learning and further model tuning. The approach is reinforcement learning from human feedback (RLHF). While few performance regressions exist on available NLP datasets, the resulting InstructGPT models have shown gains in robustness and decreased hazardous output generation.

The most significant turning point in LLM evolution marks November 30, 2022, with the release of ChatGPT (Chat generative pre-trained transformer), which was designed to help users guide a discussion toward a broad spectrum of activities like question answering, information searching, text summarizing, and more. GPT-3.5 (and eventually GPT-4), a sister model of InstructGPT taught to follow instructions in a prompt and provide a thorough answer, drives ChatGPT. The most recent and substantial LLM available from the GPT family is GPT-4 [85]. GPT-4, which debuted in March 2023, is a multi-modal LLM that can generate text outputs from images and text inputs. GPT-4 shows human-level performance on several professional and academic standards, including passing a simulated bar exam with a score in the top 10% of test takers [85], but still less adept than humans in some of the most challenging real-world circumstances. Like early GPT models, GPT-4 was initially pre-trained to predict upcoming tokens on vast text corpora, then fine-tuned by RLHF to match model behaviors with human-desired ones.

2.3.2.2 The LLaMA family Meta has released LLaMA, a compilation of open-source foundational language models that oppose GPT models. Consequently, the LLaMA family is expanding rapidly due to numerous research groups' widespread use of these models to develop more effective open-source LLMs that compete with closed-source LLMs

to develop task-specific LLMs for mission-critical applications. The initial set of LLaMA models [86] was published in February 2023, encompassing parameters ranging from 7 to 65 B. These models are pre-trained on trillions of tokens gathered from publicly available datasets. LLaMA employs the transformer architecture of GPT-3 with a few minor architectural modifications. These modifications include using a SwiGLU activation function in place of ReLU, rotary positional embeddings in place of absolute positional embedding, and root-mean-squared layer normalization in place of standard layer normalization. The proprietary GPT-3 (175B) model outperforms the open-source LLaMA-13B model on most benchmarks, rendering it an appropriate baseline for LLM research. In collaboration with Microsoft, Meta released the LLaMA-2 collection [87] in July 2023. This collection comprises both foundation language models and Chat models optimized for dialog, referred to as LLaMA-2 Chat. The LLaMA-2 Chat models were reported to outperform other open-source models on numerous public benchmarks. The procedure commences with the pre-training of LLaMA-2 using publicly available online data. Subsequently, an initial version of LLaMA-2 Chat is constructed through supervised fine-tuning. This is followed by iterative model refinement through proximal policy optimization, rejection sampling, and RLHF. To prevent the reward model from being altered excessively, which could compromise the stability of LLaMA model training, it is essential to accumulate human feedback to revise the reward model during the RLHF stage.

Using 52 K instruction-following demonstrations generated in the manner of self-instruct using GPT-3.5 (text-davinci-003), Alpaca [88] is fine-tuned from the LLaMA-7B model. Alpaca is an exceptionally cost-effective option for academic research and training. Despite its significantly reduced size, the Alpaca performs similarly to GPT-3.5 on the self-instruct evaluation set. By fine-tuning LLaMA on user-shared conversations collected from ShareGPT, the Vicuna team has developed a 13B discussion model, Vicuna 13B. Preliminary evaluations conducted with GPT 4 as an evaluator indicate that Vicuna-13B surpasses the quality of OpenAI's ChatGPT and Google's Bard by over 90% while also outperforming other models, such as LLaMA and Stanford Alpaca in over 90% of instances. Vicuna-13B also has the advantage of having a relatively low computational demand for model training. The Vicuna-13B training cost is a modest $300.

Guanaco models [89] are similarly fine-tuned LLaMA models that employ instruction-following data, as are Vicuna and Alpaca. However, the finetuning is performed with exceptional efficiency using QLoRA, allowing for the fine-tuning of a 65B parameter model on a single 48 GB GPU. Through a frozen, 4-bit quantized pre-trained language model, QLoRA back-propagates gradients into Low-Rank Adapters (LoRA). The most advanced Guanaco model surpasses all previously released models on the Vicuna benchmark, achieving a performance level of 99.3% of ChatGPT with a mere 24 h of fine-tuning on a single GPU. Koala [90] is another instruction-following language model constructed on LLaMA. However, it is distinguished by its emphasis on interaction data, which includes user inputs and responses generated by competent closed-source conversation models like ChatGPT. According to human evaluations of real-world user queries, the Koala-13B model is competitive with state-of-the-art messaging models. Mistral-7B [91] is a 7B-parameter language paradigm designed to achieve exceptional performance and efficiency. Mistral-7B surpasses the most influential open-source 13B model (LLaMA-2-13B) in all evaluated benchmarks and the most influential 34B model (LLaMA-34B) in code generation, mathematics, and reasoning. For faster inference, this model employs grouped-query attention, combined with sliding window attention, to efficiently manage sequences of any length at a reduced inference cost. The family is expanding rapidly as a result of the development of additional instruction-following models on LLaMA or LLaMA 2, such as Code LLaMA [92], Gorilla [92], Giraffe [93], Vigogne [93], Tulu 65B [93], Long LLaMA [94], and Stable Beluga2 [95].

2.3.2.3 The PaLM family Google is responsible for developing the PaLM (Pathways language model) family. The initial PaLM model [96] was disclosed in April 2022 and remained confidential until March 2023. It is an LLM that is based on a 540B parameter transformer. The model is pre-trained on a high-quality text corpus that includes 780 billion tokens and covers various natural language tasks and use cases. PaLM is pre-trained on 6144 TPU v4 processors using the pathways system, facilitating highly efficient training across numerous TPU Pods. PaLM has achieved state-of-the-art few-shot learning results on hundreds of language understanding and generation benchmarks, thereby demonstrating the ongoing benefits of scaling. PaLM 540B not only surpasses state-of-the-art fine-tuned models on a suite of multi-step reasoning tasks but also performs at a level comparable to humans on the recently released BIG-bench benchmark. The UL2R method continuously trains the U-PaLM models of 8B, 62B, and 540B scales on PaLM. This method involves continuing to train LLMs on a few steps with UL2's mixture-of-denoiser objective. A computational savings rate of approximately 2 $\times$ is reported. Flan-PaLM is subsequently instruction-fine-tuned from U-PaLM [97]. Flan-PaLM's finetuning is conducted using a significantly greater number of tasks, larger model sizes, and chain-of-thought data than other instruction finetuning work mentioned above. Consequently, Flan-PaLM substantially surpasses the performance of previous instruction-following models. For example, the Flan PaLM-540B, which is instruction-finetuned on 1.8 K tasks,

outperforms the PaLM-540B by a significant margin (+9.4% on average).

PaLM-2 [98] is a more compute-efficient LLM with superior multilingual and reasoning capabilities compared to its progenitor, PaLM. PaLM-2 is trained with a combination of objectives. PaLM-2 substantially enhances the model's performance on downstream tasks across various model sizes through comprehensive evaluations of English, multilingual, and reasoning tasks. Additionally, it demonstrates a more efficient and rapid inference process than PaLM. Med-PaLM [99] is a domain-specific PaLM designed to provide high-quality responses to medical inquiries. Instruction prompt tuning, a parameter-efficient method for aligning LLMs to novel domains using a few exemplars, is used to finetune Med-PaLM on PaLM. Although it is still inferior to human clinicians, Med-PaLM achieves highly encouraging outcomes on numerous healthcare duties. Med-PaLM 2 enhances Med PaLM by employing ensemble prompting and med-domain finetuning. Med-PaLM 2 scored 86.5% on the MedQA dataset, a benchmark combining six extant open-question–answer datasets that cover professional medical examinations, research, and consumer inquiries. This marks a significant improvement over Med-PaLM by over 19% and establishes a new state-of-the-art.

In the subsequent sections, we provide summaries of well-known pre-trained LLMs with significant discoveries that have changed the course of research and development in NLP. These LLMs have considerably improved the performance in NLU and NLG domains and are widely fine-tuned for downstream tasks.

2.3.2.4 Pre-Trained general purpose LLMs T5 [77] is an encoder-decoder model employing unified text-to-text training for all NLP problems. T5 places layer normalization outside the residual path in a conventional transformer model [62]. It uses masked language modeling as a pre-training objective where spans (consecutive tokens) are replaced with a single mask instead of separate masks for each token. This type of masking speeds up the training as it produces shorter sequences. After pre-training, the model is fine-tuned using adapter layers for downstream tasks. GPT-3 [81] has the same architecture as that of GPT-2 [5] but with dense and sparse attention in transformer layers similar to the sparse transformer [63]. It shows that large models can train on larger batch sizes with a lower learning rate; GPT-3 uses the gradient noise scale to decide the batch size during training. Overall, GPT-3 increases model parameters to 175B, showing that the performance of large language models improves with the scale and is competitive with the fine-tuned models. mT5 [78] is a multilingual T5 model trained on the mC4 dataset with 101 languages. The dataset is extracted from the public common crawl scrape. The model uses a larger vocabulary size of 250,000 to cover multiple languages. To avoid over-fitting or under-fitting for a language, mT5 employs a data sampling procedure to select samples from all languages.

PanGu-α [100] is an autoregressive model with a query layer at the end of standard transformer layers to predict the next token. Its structure is similar to the transformer layer but with an additional embedding for the following position in the attention mechanism, given in Eq. (1).

$$a = p_n W_h^q W_h^k T H_L^T \tag{1}$$

Cost-efficient Pre-trained language Models (CPM-2) [101] is a pre-trained model in bilingual (English and Chinese) on 11B and 198B mixture-of-experts (MoE) models on the Wu DaoCorpus [102] dataset. The models are trained with knowledge inheritance, starting with only the Chinese language in the first stage and then adding English and Chinese data. Moreover, to use the model for downstream tasks, CPM-2 experimented with complete fine-tuning and prompt fine-tuning, where only prompt-related parameters are updated by inserting prompts at various positions, front, middle, and back. CPM-2 also proposes INFMOE, a memory-efficient framework with a strategy to dynamically offload parameters to the CPU for inference at a 100B scale. It overlaps data movement with inference computation for lower inference time. ERNIE 3.0 [103] takes inspiration from multi-task learning to build a modular architecture using Transformer-XL as the backbone. The universal representation module is shared by all the tasks, which serve as the basic block for task-specific representation modules, which are all trained jointly for natural language understanding, natural language generation, and knowledge extraction. This LLM is primarily focused on the Chinese language, claims to train on the most significant Chinese text corpora for LLM training, and has achieved state-of-the-art performances in 54 Chinese NLP tasks.

Jurassic-1 [104] is a pair of auto-regressive language models, including a 7B-parameter J1-Large model and a 178B-parameter J1-Jumbo model. The training vocabulary of Jurassic-1 comprises word pieces, complete words, and multi-word expressions without any word boundaries, where possible out-of-vocabulary instances are interpreted as Unicode bytes. Compared to the GPT-3 counterparts, the Jurassic-1 models apply a more balanced depth-to-width self-attention architecture and an improved tokenizer for a faster prediction, achieving a comparable performance in zero-shot learning tasks and a superior performance in few-shot learning tasks. HyperCLOVA [105] is a Korean language model with GPT-3 architecture. Yuan 1.0 [106] is trained on a Chinese corpus with 5 TB of high-quality text collected from the Internet. A massive data filtering system (MDFS) built on Spark has been developed to process raw data via coarse and fine filtering techniques. To speed

Table 2 Model architecture details of Gopher with different number of parameters

Model	Layers	Number heads	key/value size	d_{model}	Max LR	Batch size (M)
44 M	8	16	32	512	6×10^{-4}	0.25
117 M	12	12	64	768	6×10^{-4}	0.25
417 M	12	12	128	1,536	2×10^{-4}	0.25
1.4B	24	16	128	2,048	2×10^{-4}	0.25
7.1B	32	32	128	4,096	1.2×10^{-4}	2
Gopher 280B	80	128	128	16,384	4×10^{-5}	3→ 6 M

up the training of Yuan 1.0 to save energy expenses and carbon emissions, various factors that improve the performance of distributed training are incorporated in architecture and training like increasing the number of hidden sizes improves pipeline and tensor parallelism performance, larger micro batches improve pipeline parallelism performance, and more significant global batch size improves data parallelism performance. The Yuan 1.0 model performs well on text classification, winograd schema, natural language inference, and reading comprehension tasks. The Gopher family of models ranges from 44 M to 280B parameters in size to study the effect of scale on the LLM's performance [107]. The 280B model beats GPT 3 [6], Jurrasic-1 [107], MT-NLG [107], and others on 81% of the evaluated tasks. These models were evaluated on 152 diverse tasks, achieving state-of-the-art performance across the majority. The number of layers, the key/value size, and other hyper-parameters of different model sizes are shown in Table 2.

By training a bigger model with 26 times the parameters of the latter, ERNIE 3.0 Titan [108] expands upon ERNIE 3.0. With its larger model, it beat other top-tier models in 68 natural language processing tasks. As part of its multi-task learning setup, ERNIE 3.0 Titan incorporates a new task—credible and controlled generations—to ensure factual consistency in the produced text. Including adjustable language modeling losses and self-supervised adversarial losses in the pre-training process gives ERNIE 3.0 Titan an advantage over other LLMs in assessing factual QA task sets that were manually chosen. It combines an auto-regressive network with an auto-encoding network, making it easy to train a model for generating or interpreting natural language via zero-shot learning, few-shot learning, or fine-tuning. Figure 9 shows the Ernie 3.0 model's architecture.

An auto-regressive model called GPT-NeoX-20B [109] was trained on the Pile dataset without data deduplication. It is based on GPT-3 with a few architectural design variations, like parallel attention and feed-forward layers in a transformer block, that increase throughput by 15%. Although it only applies rotating positional embedding to a quarter of the embedding vector dimension, it is effective nevertheless. While GPT-3 uses dense and sparse layers, GPT-NeoX-20B only employs dense layers. The model selects its hyperparameters utilizing the approach and uses interpolation to get values between 13 and 175 B models for the 20B model since hyperparameter tuning at this scale is challenging. Using tensor and pipeline parallelism, the training of the model is divided across GPUs.

$$x + Attn(LN_1(x) + FF(LN_2(x)) \tag{2}$$

Open Pre-trained Transformers (OPT) [110], a suite of decoder-only pre-trained transformers, is a clone of GPT-3, developed to open-source a model that replicates GPT-3 performance. Training of OPT employs dynamic loss scaling and restarts from an earlier checkpoint with a lower learning rate whenever loss divergence is observed. Overall, the performance of OPT-175B models is comparable to the GPT3-175B model. The OPT models' parameters are shown in Table 3.

BLOOM [111] is a causal open-source LLM decoder-only transformer language model trained on ROOTS corpus, a dataset comprising hundreds of sources in 46 natural and 13 programming languages (59 in total). Its architecture is differentiated into ALiBi positional embedding and an additional normalization layer after the embedding layer, as suggested by the bitsandbytes1 library. These changes stabilize training with improved downstream performance. It is a 176B parameter language model designed and built thanks to the collaboration of hundreds of researchers [112]. An overview of BLOOM architecture is shown in Fig. 10.

The Generalist Language Model (GLaM) [113] represents a family of language models using a sparsely activated decoder-only mixture-of-experts (MoE) structure. The experts are sparsely activated to gain more model capacity while reducing complexity, and only the best two experts are used to process each input token. The largest GLaM model, GLaM (64B/64E), is about 7 × larger than GPT 3, while only a part of the parameters is activated per input token. The largest GLaM (64B/64E) model achieves better overall results than GPT-3 while consuming only one-third of GPT-3's training energy. MT-NLG [114] is a 530B causal decoder based on GPT-2 architecture, roughly 3 × GPT-3 model

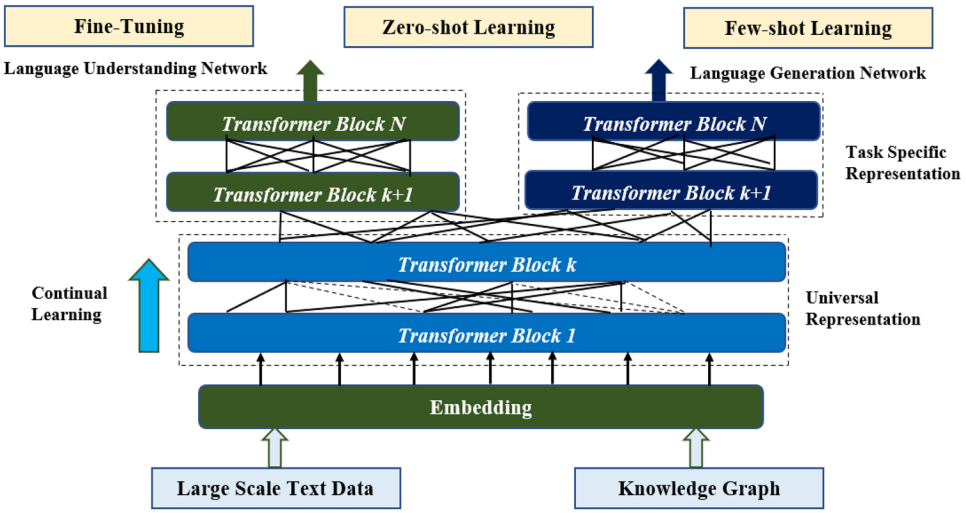

Fig. 9 High-level model architecture of ERNIE 3.0

Table 3 Different OPT models' architecture details

Model	#L	#H	d_{model}	LR	Batch (M)
125 M	12	12	768	6.0e−4	0.5
350 M	24	16	1024	3.0e−4	0.5
1.3B	24	32	2048	2.0e−4	1
2.7B	32	32	2560	1.0e−4	1
6.7B	32	32	4096	1.0e−4	2
13B	40	40	5120	1.0e−4	4
30B	48	56	7168	1.0e−4	4
66B	64	72	9216	0.8e−4	2
175B	96	96	12,288	1.2e−4	2

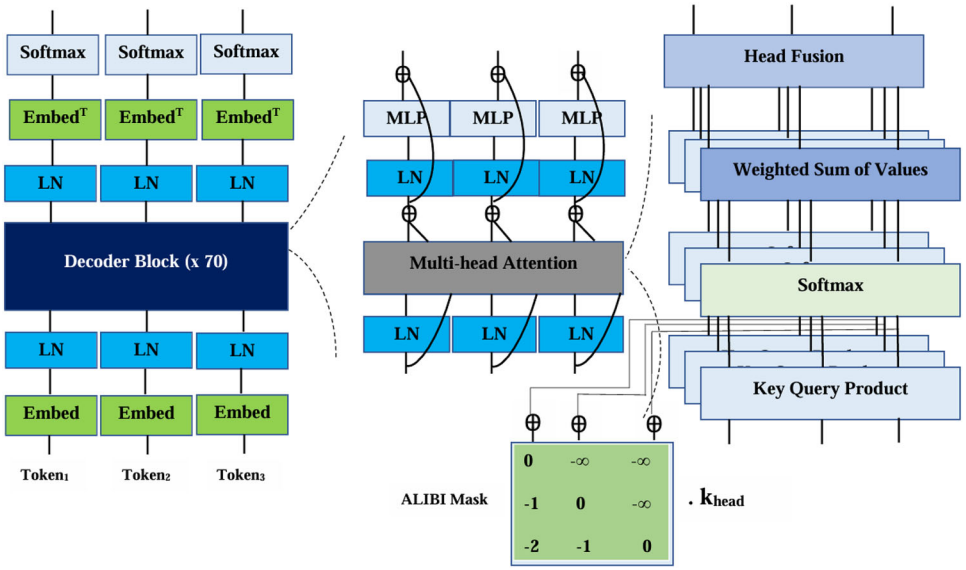

Fig. 10 An overview of BLOOM architecture

parameters. It is trained on filtered, high-quality data collected from various public datasets. It blends multiple types of datasets in a single batch, which beats GPT-3 on several evaluations. Chinchilla [115] is a causal decoder trained on the same dataset as the Gopher but with a slightly different data sampling distribution. The model architecture is similar to the one used for Gopher, except for the AdamW optimizer instead of Adam. It identifies the relationship that model size should be doubled for every doubling of training tokens. Over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens are trained to get the estimates for compute-optimal training under a given budget. AlexaTM [116] is an encoder-decoder model where encoder weights and decoder embeddings are initialized with a pre-trained encoder to speed up training. The encoder stays frozen for the initial 100 k steps and later unfreezed for end-to-end training. The model is trained on denoising and causal language modeling (CLM) objectives, concatenating tokens at the beginning for mode switching. During training, the CLM task is applied 20% of the time, which improves the in-context learning performance.

PaLM [96] is a causal decoder with parallel attention and feed-forward layers similar to Eq. (2), speeding up training 15 times faster. Additional conventional transformer model changes include SwiGLU activation, RoPE embeddings, multi-query attention that saves computation cost during decoding, and shared input–output embeddings. U-PaLM [97] is a method that trains PaLM for 0.1% additional compute with UL2 (also named as UL2Restore) objective using the same dataset and outperforms baseline significantly on various NLP tasks, including zero-shot, few-shot, common sense reasoning, CoT, etc. Training with UL2R involves converting a causal decoder PaLM to a non-causal decoder PaLM and employing 50% sequential denoising, 25% regular denoising, and 25% extreme denoising loss functions. PaLM-2 [98] is a smaller multi-lingual variant of PaLM, trained for more significant iterations on a better-quality dataset. The PaLM-2 significantly improves over PaLM while reducing training and inference costs due to its smaller size. To lessen toxicity and memorization, it appends unique tokens with a fraction of pre-training data, which shows a reduction in generating harmful responses. UL2 [117] is an encoder-decoder architecture trained using a mixture of denoisers (MoD) objectives. Denoisers include (1) R-Denoiser: a regular span masking; (2) S-Denoiser: which corrupts consecutive tokens of an extensive sequence; and (3) X-Denoiser: which corrupts many tokens randomly. UL2 includes a denoiser token from R, S, and X during pre-training to represent a denoising setup. It helps improve the fine-tuning performance for downstream tasks that bind the task to one of the upstream training modes.

GLM-130B [118] is a bilingual (English and Chinese) model trained using an auto-regressive mask infilling pre-training objective similar to the GLM. This training style makes the model bidirectional as compared PREPRINT 11 to GPT-3, which is unidirectional. Opposite to the GLM, the training of GLM-130B includes a small amount of multi-task instruction pre-training data (5% of the total data) along with the self-supervised mask infilling. To stabilize the training, embedding layer gradient shrink is applied. LLaMA [86] is a set of decoder-only language models varying from 7 to 70B parameters. LLaMA models series is the most famous among the community for parameter efficiency and instruction tuning. LLaMA-1 [86] Implements efficient causal attention by not storing and computing masked attention weights and key/query scores. Another optimization is reducing the number of activations recomputed in the backward pass. LLaMA-2 [87] focuses more on fine-tuning a safer and better LLaMA-2-Chat model for dialogue generation. The pre-trained model has 40% more training data with a more significant context length and grouped-query attention. PanGu-Σ [119] us an autoregressive model with parameters copied from PanGu-α and extended to a trillion scale with random routed experts (RRE). RRE is similar to the MoE architecture, with distinctions at the second level, where tokens are randomly routed to experts in a domain instead of using a learnable gating method. The model has bottom layers densely activated and shared across all domains, whereas top layers are sparsely activated according to the domain. This training style allows for extracting task-specific models and reduces catastrophic forgetting effects in the case of continual learning.

2.3.2.5 Pre-trained coding-based LLMs CodeGen [120] and the PaLM [96] are architecturally comparable; both use RoPE embeddings, MLP layers, and parallel attention. The model is trained using data from programming languages and natural language consecutively on the given datasets: (1) PILE, (2) BIGQUERY, and (3) BIGPYTHON. The idea is to make it easier to produce lengthy code sequences by passing in the output of the previous question and the code it generated as input to the following prompt. CodeGen has made an open-source Multi-Turn Programming Benchmark (MTPB) available to evaluate multi-step program synthesis. Codex [121] is an LLM trained on a subset of public Python Github repositories to generate code from docstrings. Computer programming is an iterative process in which the programs are often debugged and updated before fulfilling the requirements. Similarly, Codex generates 100 program versions by repetitive sampling for a given description, producing a working solution for 77.5% of the problems passing unit tests. Its mighty version powers Github Copilot2. AlphaCode

[122] collects 300 M–41B-parameter language models for competition-level code creation. Memory and cache costs are reduced via multi-query attention. AlphaCode models are pre-trained on filtered GitHub code in common languages and fine-tuned on CodeContests, a new competitive programming dataset, since competitive programming challenges need deep thinking and a mastery of complicated natural language techniques. Problems, solutions, and test cases from Codeforces dominate the CodeContests dataset3. Standard language modeling goals are used for pre-training, and GOLD [123] with tempering is used for CodeContests data fine-tuning. AlphaCode performs well in Codeforces-hosted simulated programming contests, ranking 54.3% among over 5000 contestants and 28% of recently participating users.

CodeT5 + [124] is based on CodeT5, with a shallow encoder and deep decoder, trained in multiple stages, initially unimodal data (code) and later bimodal data (text-code pairs). Each training stage has different objectives and activates different model blocks, such as an encoder, decoder, or both, according to the task. The unimodal pre-training includes span denoising and CLM objectives, whereas bimodal pre-training objectives contain contrastive learning, matching, and CLM for text-code pairs. It adds unique tokens with the text to enable task modes, for example, [CLS] for contrastive loss, [Match] for text-code matching, etc. StarCoder [125] is a decoder-only model with SantaCoder architecture, employing Flash attention to scale the context length to 8 k. The StarCoder trains an encoder to filter names, emails, and other personal data from the training data. Its fine-tuned variant outperforms PaLM, LLaMA, and LAMDA on HumanEval and MBPP benchmarks.

2.3.2.6 Pre-trained large language models for scientific knowledge Galactica [126] is a large, curated corpus of human scientific knowledge with 48 million papers, textbooks, lecture.

notes, millions of compounds and proteins, scientific websites, encyclopedias, and more are trained using metaseq library3, built on PyTorch and fair scale. The model wraps reasoning datasets with < work > token to provide step-by-step reasoning context to the model, which has been shown to improve the performance on reasoning tasks. LaMDA [127] is a decoder-only model trained on public dialog data, public dialog utterances, and public online publications, with more than 90% of the pre-training data in English. LaMDA is trained with the goal of providing answers that are high-quality, safe, and grounded. To do this, discriminative and generative fine-tuning approaches are used to improve the model's safety and quality characteristics. Consequently, the LaMDA models may be used as a generic language model for various activities. BloombergGPT [128] is a non-causal decoder model trained using both financial ("FINPILE" from the Bloomberg archive) and general-purpose datasets. The model's architecture is similar to the BLOOM [111] and OPT [110]. It allocates 50B parameters to different blocks of the model using the approach. For practical training, BloombergGPT packs documents with <|endoftext|> to use maximum sequence length, warmup batch size from 1024 to 2048, and manually reduce the learning rate multiple times during the training. Xuan Yuan 2.0 [129] is a Chinese financial chat model with BLOOM's [111] architecture trained on a combination of general purpose, financial, general purpose instructions, and financial institutions datasets. It combined the pre-training and fine-tuning stages to avoid catastrophic forgetting.

2.3.2.7 Other representative of large language models In addition to the models discussed in the previous subsections, other popular LLMs do not belong to the categories mentioned earlier, yet they have achieved excellent performance and have pushed the LLM field forward. We briefly describe these LLMs in this subsection. FLAN [130] is a simple instruction-tuned model that came into the picture by improving the zero-shot learning abilities of language models. It showed that instruction-tuning language models on a collection of datasets described via instructions substantially improves zero-shot performance on unseen tasks. It was evaluated on a 137B parameter pre-trained language model, and instruction tuned it on over 60 NLP datasets verbalized via natural language instruction templates. Figure 11 compares instruction tuning with pretrain–finetune and prompting.

T0 [131] is a system for efficiently mapping natural language tasks into a human-readable prompted form. A large set of supervised datasets was converted, each with multiple prompts with diverse wording. These prompted datasets allow for benchmarking the ability of a model to perform completely held-out tasks. Then, a T0 encoder-decoder model is developed to consume textual inputs and produce target responses. The model is trained on a multitask mixture of NLP datasets partitioned into different tasks. RETRO [132] is an enhanced auto-regressive language model by conditioning on document chunks retrieved from a large corpus based on local similarity with preceding tokens. Using a 2-trillion-token database, the retrieval-enhanced transformer (RETRO) obtains a performance comparable to GPT-3 and Jurassic-1 [83] on the Pile, despite using 25% fewer parameters. It combines a frozen Bert retriever, a differentiable encoder, and a chunked cross-attention mechanism to predict tokens based on an order of magnitude more data than what is typically consumed during training. Sparrow [133] is an information-seeking dialogue agent trained to be more helpful, correct, and harmless than prompted language model baselines. It was trained from reinforcement learning from human feedback with two new additions to help human raters judge agent behavior. Minerva [134] is a large language model pre-trained on general natural language data

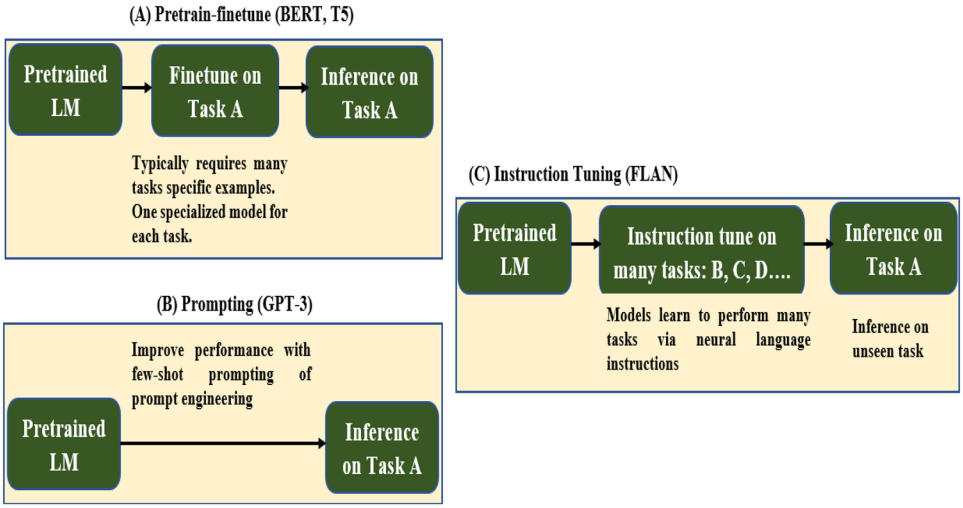

Fig. 11 Comparison of instruction tuning with pre-train finetune and prompting

and further trained on technical content to tackle previous LLM struggles with quantitative reasoning, such as solving mathematics, science, and engineering problems.

MoD [135] is a generalized and unified perspective for self-supervision in NLP. It shows how different pre-training objectives can be cast as one another and how interpolating between different objectives can be effective. It employed a mixture-of-denoisers (MoD), a pre-training objective that combines diverse pre-training paradigms. Pythia [136] is a suite of 16 LLMs trained on public data seen in the same order and ranging in size from 70 M to 12B parameters. Orca [137] is a 13-billion parameter model that learns to imitate the reasoning process of large foundation models. Orca learns from rich signals from GPT-4, including explanation traces, step-by-step thought processes, and other complex instructions guided by teacher assistance from ChatGPT. KOSMOS [138] is a multimodal large language model (MLLM) that can perceive general modalities, learn in context (few-shot), and follow instructions (zero-shot). Specifically, it was trained from scratch on web-scale multimodal corpora, including arbitrarily interleaved text and images, image-caption pairs, and text data. Experimental results show that KOSMOS 1 achieves impressive performance on (i) Language understanding, generation, and even OCR-free NLP (directly fed with document images), (ii) Perception-language tasks, including multimodal dialogue, image captioning, visual question answering, and (iii) Vision tasks, such as image recognition with descriptions (specifying classification via text instructions). Gemini [139] is a new family of multimodal models exhibiting promising capabilities across image, audio, video, and text understanding. Gemini family includes three versions: Ultra for highly complex tasks, pro for enhanced performance and deployability at scale, and nano for on-device applications. Gemini architecture is built on top of transformer decoders and is trained to support 32 k context length (using efficient attention mechanisms).

Table 4 Some popular large language model frameworks

S. no	Name	S. no	Name
1	Inner Monologue	14	Gorilla
2	Megatron-Turing NLG	15	PAL
3	LongFormer	16	Claude
4	OPT-IML	17	CodeGen 2
5	MeTaLM	18	Zephyr
6	Dromedary	19	Grok
7	Palmyra	20	Qwen
8	Camel	21	Mamba
9	Yalm	22	Mixtral-8 × 7B
10	MPT	23	DocLLM
11	ORCA 2	24	DeepSeek-Coder
12	FuseLLM-7	25	LLaMA-Pro-8B
13	TinyLlama-1.1B		

Some of the other popular LLM frameworks (or techniques used for the efficient development of LLMs are given in Table 4.

Figure 12 provides an overview of some of the most representative large language model frameworks and the relevant works that have contributed to the success of large language models and helped to push the limits of large language models.

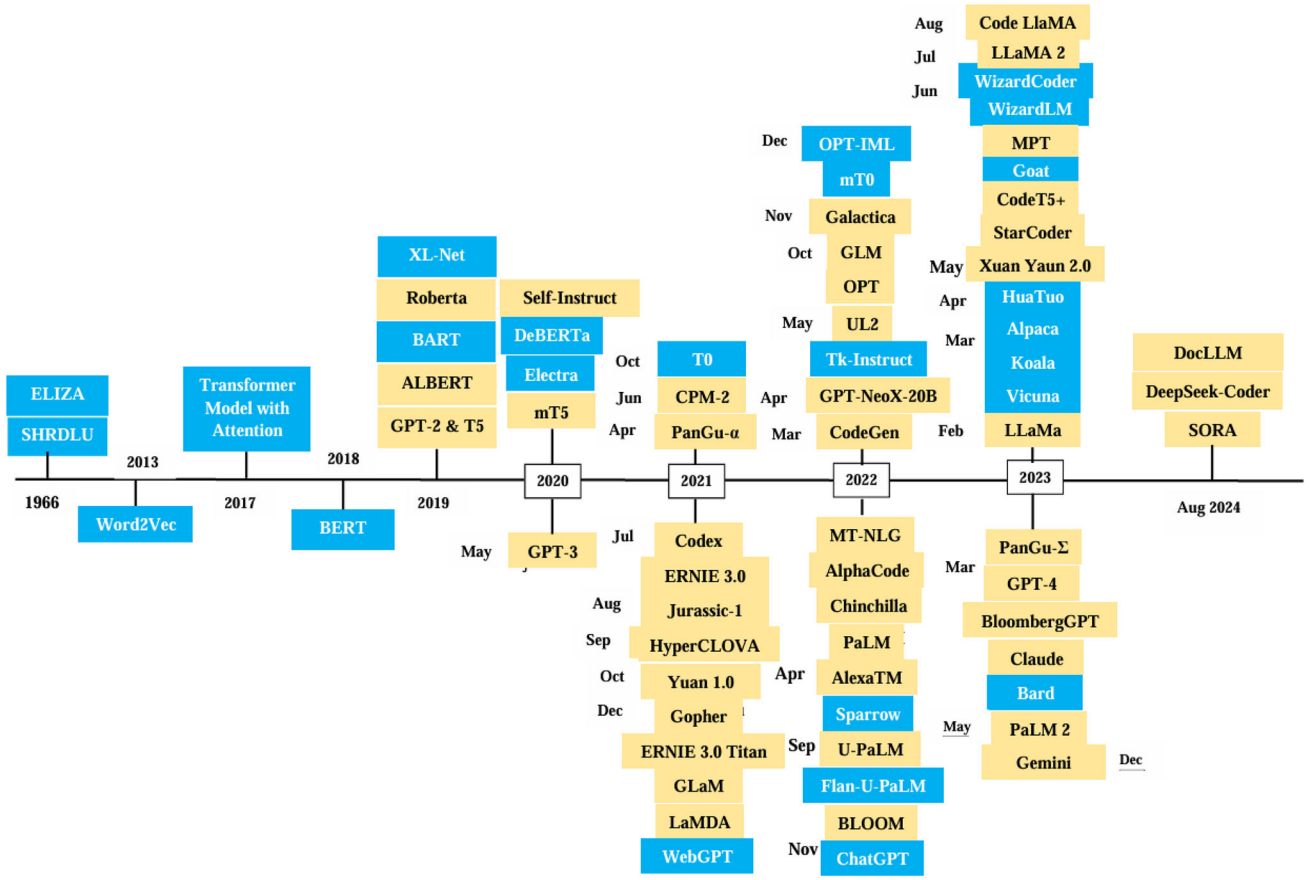

Fig. 12 Chronological display of large language model releases: light blue rectangles represent 'pre-trained' models, while yellow rectangles correspond to 'instruction-tuned' models. Models on the upper half signify open-source availability, whereas those on the bottom half are closed-source

3 Summary

Based on the discussions mentioned above, a high-level overview of all popular large language models discussed in the above sections is presented in Table 5 below.

From the above discussion, we have understood that different LLMs may be evaluated using a variety of approaches. For example, an LLM with much fewer parameters is not equivalent to one with a more significant number of parameters [140]. From this viewpoint, the research community has divided LLMs into four groups depending on their dimensional spaces: small (less than or equal to 1 billion parameters), medium (between 1 and 10 billion), large (between 10 and 100 billion), and very large (more than 100 billion). Table 7 displays all of the many classifications and applications of LLMs. Similarly, LLMs can be classified based on their primary use cases, with each LLM being either a foundation model (pre-trained language model with no instruction fine-tuning and chat fine-tuning), an instruction model (pre-trained language model with only instruction fine-tuning), or a chat model (pre-trained language model with instruction and chat fine-tuning). Aside from the above categories, another category is necessary to differentiate between original and adjusted models. Original models have been provided as foundation models or fine-tuned versions. Tuned models took the initial model and refined it using other datasets or training methodologies. It's also worth noting that initial models are often foundation models that have been fine-tuned for certain datasets or even various techniques. Regardless of license, the availability of model weights is another factor in LLM categorization. Models with publicly accessible weights (even upon request) are labeled Public models, while others are labeled Private (Table 6).

4 Applications: harnessing the large language models for smart agriculture and precision farming

The application of large language models has become a hot topic in AI-related research communities and industries, with many emerging uses being discovered and explored

Table 5 A High-level overview of popular large language models

Type	Model name	#Parameters	Release	Base models	Open source	#Tokens	Training dataset
Encoder-only	BERT	110 M, 340 M	2018	–	√	137B	Books Corpus, English Wikipedia
	RoBERTa	335 M	2019	–	√	2.2 T	Books Corpus, English Wikipedia, CC-NEWS, STORIES (a subset of Common Crawl), Reddit
	ALBERT	12 M, 18 M, 60 M, 235 M	2019	–	√	137B	Books Corpus, English Wikipedia
	DeBERTa	–	2020	–	√	–	Books Corpus, English Wikipedia, STORIES, Reddit content
	XLNet	110 M, 340 M	2019	–	√	32.89B	Books Corpus, English Wikipedia, Giga5, Common Crawl, ClueWeb2012-B
Decoder-only	GPT-1	120 M	2018	–	√	1.3B	Books Corpus
	GPT-2	1.5B	2019	–	√	10B	Reddit outbound
Encoder-decoder	T5 (Base)	223 M	2019	–	√	156B	Common Crawl
	MT5 (Base)	300 M	2020	–	√	–	New Common Crawl-based dataset in 101 languages (m Common Crawl)
	BART (Base)	1.39 M	2019	–	√	–	Corrupting Text
GPT family	GPT-3	125 M, 350 M, 760 M, 1.3B, 2.7B	2020	–	×	300B	Common Crawl (filtered), WebText2, Books1, Books2, Wikipedia
	CODEX	12B	2021	GPT	√	–	Public GitHub software repositories
	WebGPT	760 M, 13B, 175B	2021	GPT-3	×	–	ELI5
	GPT-4	1.76 T	2023	–	×	13 T	-
LLaMA family	LLaMA1	7B, 13B, 33B, 65B	2023	–	√	1 T, 1.4 T	Online Sources
	LLaMA2	7B, 13B, 34B, 70B	2023	–	√	2 T	Online Sources
	Alpaca	7B	2023	LLaMA1	√	–	GPT-3.5
	Vicuna-13B	13B	2023	LLaMA1	√	–	GPT-3.5

Table 5 (continued)

Type	Model name	#Parameters	Release	Base models	Open source	#Tokens	Training dataset
	Koala	13B	2023	LLaMA	√	–	Dialogue Data
	Mistral-7B	7.3B	2023	–	√	–	–
	Code Llama	34B	2023	LLaMA2	√	500B	Publicly Available Code
	Long LLaMA	3B, 7B	2023	OpenLLaMA	√	1 T	–
	LLaMA-Pro-8B	8.3B	2024	LLaMA2-7B	√	80B	Code & Math Corpora
	TinyLlama-1.1B	1.1B	2024	LLaMA1.1B	√	3 T	SlimPajama, Starcoderdata
PaLM Family	PaLM	8B, 62B, 540B	2022	–	×	780B	Web documents, Books, Wikipedia, Conversations, GitHub code
	U-PaLM	8B, 62B, 540B	2022	–	×	1.3B	Web documents, Books, Wikipedia, Conversations, GitHub code
	PaLM-2	340B	2023	–	√	3.6 T	Web documents, Books, Code, Mathematics, Conversational data
	Med-PaLM	540B	2022	PaLM	×	780B	HealthSearchQA, MedicationQA, LiveQA
	Med-PaLM 2	-	2023	PaLM 2	×	–	MedQA, MedMCQA, HealthSearchQA, LiveQA, MedicationQA
Other popular LLMs	FLAN	137B	2021	LaMDA-PT	√	–	Web documents, Code, Dialogdata, Wikipedia
	Gopher	280B	2021	–	×	300B	Massive Text
	ERNIE 4.0	10B	2023	–	×	4 T	Chinese Text
	Retro	7.5B	2021	–	×	600B	Massive Text
	LaMDA	137B	2022	–	×	168B	Public Dialogue Data & Web Documents
	ChinChilia	70B	2022	–	×	1.4 T	Massive Text
	Galactia-120B	120B	2022	–	–	450B	–
	CodeGen	16.1B	2022	–	√	-	THE PILE, BIG QUERY, BIG PYTHON
	BLOOM	176B	2022	–	√	366B	ROOTS
	Zephyr	7.24B	2023	Mistral-7B	√	800B	Synthetic Data
	Grok-0	33B	2023	–	×	–	Online Source

Table 5 (continued)

Type	Model name	#Parameters	Release	Base models	Open source	#Tokens	Training dataset
	ORCA-2	13B	2023	LLaMA2	–	200B	–
	StartCoder	15.5B	2023	–	√	35B	GitHub
	MPT	7B	2023	–	√	1 T	RedPajama, m Common Crawl, S2ORC, Common Crawl
	Mixtral-8 × 7B	46.7B	2023	–	√	–	Instruction dataset
	Falcon 180B	180B	2023	–	√	3.5 T	RefinedWeb
	Gemini	1.8B, 3.25B	2023	–	√	–	Web documents, books, and code, image data, audio data, video data
	DeepSeek-Coder	1.3B, 6.7B, 33B	2024	–	√	2 T	GitHub's Markdown and StackExchange
	DocLLM	1B, 7B	2024	–	×	2 T	IIT-CDIPTest Collection 1.0, DocBank

Table 6 Large language model categories in the literature

Classification	Category	Description
Size	Small	Number of parameters $\leq$ 1B
	Medium	1B < Number of parameters $\leq$ 10B
	Large	10B < Number of parameters $\leq$ 100B
	Very large	100B < Number of parameters
Type	Foundation model	Pretrained language model
	Instruction model	Pretrained and instruction fine-tuned language model
	Chat model	Pretrained, instruction fine-tuned, and chat fine-tuned language model
Origin	Original model	An original model released with either a Foundation, Instruction, or Chat model
	Tuned model	A fine-tuned version of an original model
Availability	Publicly available	Model and weights are available due to request or without request
	Publicly unavailable	Model and weights are not publicly available

daily. These models, capable of understanding and generating human-like text, have found meaningful applications across various fields. This section will overview LLMs' applications in smart and precise agriculture, medicine, education, science, mathematics, law, finance, and coding. In the case of the agriculture domain, we further categorized the applications into four more domains, ranging from large language models in agricultural applications, large vision models in agrarian applications, multimodal large language models and model assessment, and reinforcement learning foundation models (RLFMs) in agriculture. These four subcategorizations are further divided into various sub-sections. While each of these domains poses different challenges, LLMs have opportunities to make significant contributions given their wide-ranging scope of applicability.

4.1 Large language models in agricultural applications

Large language models are also known as large foundation models (LFM), and in agriculture, we will be focussing more or less on agricultural foundation models (AFMs). Here, we present current applications of these models in the realm of intelligent agriculture, followed by a discussion on emerging areas where these foundation models are most applicable.

Large-scale language models (LLMs) have revolutionized design, development, and deployment domains by facilitating interactions reminiscent of human communication. Complete computer-aided diagnostics (CAD) models, code evaluation, or autonomous robot fabrication are all tasks that LLMs cannot do now. Agricultural researchers and farmers, for instance, depend on massive models to carry out several critical tasks when given an image of a soybean field. The first step is for the considerable model to detect any unusual signs on soybean leaves, such as water spots. After that, it must determine what is wrong with plants, such as bacterial wilt disease. Pseudorabies is an example of an underlying ailment that the model must next identify. Lastly, a bactericide spray or other suitable treatment plan has to be devised. Interestingly, these LLMs are finding more and more uses in the research and development process for agricultural applications. Stella et al. [141] explicitly addressed the inclusion of large language models in the robotic system design process. In particular, they show how to make a robotic gripper that picks tomatoes well. While still brainstorming, researchers look to LLMs like ChatGPT (OpenAI, 2022) for insight into the field's possible obstacles and prospects. They use this data to choose the most exciting and promising routes, then use more communication with the LLM to limit the design options. Collaboration in this context encompasses not just conceptualization but also technical execution across various disciplines of expertise and abstraction levels. Using the AI's knowledge, the human collaborator can access information beyond their competence throughout this process. These broad strokes must be transformed into a real, fully operational robot in the next, more technically oriented stage of the design process.

Many conversation and question-answering (QA) tools are made to help people with trouble thinking [142]. For example, a robot built on a recurrent neural network (RNN) is intended to answer questions about checking soil, protecting plants, and managing nutrients [142]. These QA and conversation systems and robots can answer most questions correctly without human help. Still, they aren't very good at solving complicated problems because their models are too small and lack training data. So, the farming field needs large models to help the growth of QA systems, conversation systems, and robots. AI studies have recently made progress, though, which shows that these algorithms can help a lot with running software, thinking mathematically [143], and even making shapes [143]. We believe that AI methods will be able to handle these jobs in the future, but for now, the actual application is still done by people and AI models working together. Ultimately, the person tweaked the code that the LLM offered, finished the CAD design, and watched the robot's construction. Lu et al. [144] suggested how LLMs could be used to organize unorganized metadata, change metadata formats, and find mistakes that might have happened during the data collection process. They also think that the next generation of LLMs will be powerful data display tools that help academics get helpful information from vast amounts of phenotype data.

4.1.1 The role of large language models in processing and generating agricultural data, providing insights and decision-making support

LLM can play many roles in the agricultural domain, such as processing and generating agricultural data, providing insights into agricultural production work, and supporting agrarian decision-making for farmers.

4.1.1.1 Large language models information extraction LLMs can take random farming text data and turn it into organized information. To begin, the text is broken up into small tokens, which are then stored as a number vector known as a word embedding. Then, LLMs look at each token's surrounding context to determine what the sentence means. These also find and group named things in the text, such as names of people, places, businesses, or specific farming terms. Lastly, LLMs use methods such as information extraction to find and get structured information from unorganized text. This can include finding the connections between entities, getting key facts, or filling out knowledge graphs. LLMs use a method called NLP to get information from data. Peng et al. [145] used LLM to automatically pull out entities and characteristics from unlabeled farm data and turn it into structured data. It helps LLM understand the general sense of the text better by extracting information from it.

4.1.1.2 Large language models data generation Generative artificial intelligence (Gen AI) models are multimodal, large language models. Insufficient training data and labels are challenging when applying specialized computer vision algorithms to agricultural vision data [146]. Furthermore, gathering data capturing the vast variances of seasonal and weather changes is tricky. Acquiring high-quality data takes a long time, and categorizing them is considerably more expensive. To solve these issues, one strategy is to fine-tune multimodal generative LLMs in the agricultural data domain. This enables the models to create extensive training data and labels, resulting in an expanded training set that closely reflects the original data distribution.

Furthermore, language-based generation models may produce images and films [147] of particular situations from written descriptions, augmenting training datasets lacking visual material. These models may also make several versions of the primary data for specific characteristics: multimodal generative LLMs can change the image's time from day to night or the weather conditions from rainy to sunny.

This helps to increase the training data and enhance the performance of downstream models.

4.1.1.3 Large language models provide insights Through textual data analysis, LLMs may identify patterns in customer preferences, market dynamics, agricultural practices, and policy advancements. These algorithms may provide insightful information on price patterns and market dynamics by examining agricultural text data from news stories, reports, and social media sources. This helps farmers learn about areas that are not related to agriculture. Several studies show a discernible increase in incorporating LLMs into various design and development phases for agricultural applications [148]. LLMs may assist researchers in lowering the chance of failure by providing their opinions on whether the technology is practical, even while they do not give full technical support. LLMs cannot autonomously manufacture robots, assess programming, or create detailed CAD models. However, new developments in LLM research indicate that these algorithms may greatly help with software execution, mathematical reasoning, and even form generation [149]. Lu et al. [144] employed LLMs to organize unstructured information, facilitate metadata translation across formats, and identify possible mistakes in the data-collecting process in the specific areas of attention. Additionally, they saw the next generation of LLMs as compelling data visualization tools, and they expected that these sophisticated models would greatly assist researchers, allowing them to glean valuable insights from vast amounts of phenotypic data. It is vital to remember that while LLMs provide insights that might indirectly assist farmers in solving a limited number of agricultural problems, their usage should be accompanied by human judgment and domain experience. In other words, human experience and the insights LLMs offer are inextricably linked.

4.1.1.4 Large language models provide decision-making support for farmers New research claims that ChatGPT can understand natural language requests, extract useful textual and visual information, choose relevant language and vision activities, and successfully convey the outcomes to people. To tackle AI challenges, Shen et al. [150] suggested a HuggingGPT system that uses a language interface to link LLM with AI models based on HuggingFace. According to this, LLMs may act as the central decision-making component, directing other AI models so that the system as a whole can do the jobs that farmers have suggested. LLM may be used in agriculture as the foundation of decision-making to assist in resolving the issues raised by farmers [150]. To handle challenging AI tasks, LLMs may be able to act as controllers, monitoring and controlling the activities of current AI models. An uneducated farmer supplied a picture and utilized voice to check the system. Upon receiving a job request, LLM first breaks the work into smaller tasks and chooses the best AI model for each task, depending on its requirements. For example, converting farmers' audio into text requires an audio-to-text model (Amazon transcribe, Whisper) [151]. It is also necessary to recognize the sent image and integrate the text from the audio conversion in the previous step to get a text response (vit-gpt2). Considering that the farmer is illiterate, it is necessary to further convert the text response into audio and ultimately get the audio response (Fastspeech) [152]. Although LLM does not play a role in solving problems throughout the system, as a "conductor," it can coordinate various AI models to complete subtasks, thereby gradually solving complex tasks and playing a core role in decision-making support.

4.1.2 Smart farming and livestock management

Smart farming is a method of agricultural management that optimizes the amount of human labor required while increasing the quantity and quality of products through modern information and communication technologies. Modern producers can access various technologies, including sensors, software, connectivity, location, robotics, and data analytics. Some of the critical domains in intelligent farming and livestock management using LLMs are discussed in this section.

4.1.2.1 Large language models in smart crop management Smart crop management [153] has emerged as a new study area for sustainable farming. It incorporates cutting-edge techniques from big data, IoT, and artificial intelligence into crop management activities, including harvesting, disease detection, monitoring plant development, detecting diseases, monitoring yield, harvesting plant growth, and yield monitoring [154]. However, creating such algorithms for precision crop management is complex and complicated, especially considering variables like various crop types, changing climatic conditions, and few or difficult-to-get training datasets. Additionally, these systems often need a variety of data inputs from different disciplines. For instance, a deep learning technique that uses satellite images to evaluate crop health may need market data to forecast the best time to harvest based on price patterns and soil sensor data to comprehend subsurface moisture levels. Deep research insights are necessary for designing such algorithms, making integrating information from unrelated domains challenging. Given the above-noted difficulties, large language models provide a viable remedy. These models need little fine-tuning for other domains since they were trained on datasets from several fields. They have a lot of promise as instruments for astute crop management. For example, LLMs facilitate software development and design by enabling academics to access information outside their area of expertise [141].

4.1.2.2 Large language models in smart livestock farming

"Smart livestock farming," a new automated approach, incorporates cutting-edge technology such as deep learning, the Internet of Things, and intelligent control systems for livestock production [155]. This includes automating crucial processes like egg collection, animal monitoring, and precise environmental conditions. The primary objective of this innovative strategy is to significantly lower material, labor, and maintenance costs while improving operating efficiency on cattle ranches. Despite its innovative and promising nature, smart live animal husbandry has its share of challenges. It takes knowledge of various disciplines, including data analytics, advanced technology, animal science, and environmental research, to develop and implement such an AI system. The core of this integrated approach is the processing and management of vast volumes of different data from several sources. Utilizing cutting-edge data analytics and artificial intelligence methods and incorporating expertise from other disciplines is essential to overcoming these challenges. Despite these challenges, LLMs can make smart live animal husbandry a reality [156]. These comprehensive models, created on large-scale multimodal data and including knowledge from other fields, can effectively handle the complexities of multi-dimensional information [157]. Because of their ability to learn from and provide insights from large datasets, LLMs have the potential to make a substantial contribution to the creation of intelligent and effective livestock farming systems.

Animal farmers have lately extended their use of computer vision systems aided by deep learning algorithms for phenotyping and behavioral identification and monitoring at the group or individual level to improve production efficiency and animal welfare. One potential stumbling block to deep learning use is the laborious process of gathering and labeling massive volumes of animal image data. To produce high-quality images of dairy goats, Li and Tang [158] tested a proposal that used a normalized self-attention mechanism and substituted multi-class labels (colors, backgrounds, and behaviors) for one-hot labels. Using images of muzzle patterns as input, Singh et al. [159] used an LLM model to produce better images for cow identification. To train the model to map out image improvements, authors utilized low-quality samples by intentionally degrading the originals and binarized muzzle pattern images. Using synthetic images, feature extraction, and matching, four different cow breeds were positively identified with a 98.86% success rate. A critical omission from their research is mentioning how well the initial images identified the livestock. The use of wearable sensors for ambulatory tracking and behavior monitoring has been the subject of much study. Ahmed et al. [160] used activity sensors attached to chicks to assess time-series data to distinguish between healthy and unwell birds.

4.1.2.3 Large language models in smart aquaculture farming

Precision aquaculture farming aims to maximize production efficiency, environmental sustainability, and animal welfare. These systems typically consist of hundreds of linked sensors that gather information on fish tagging, temperature, oxygen levels, weather, chlorophyll, images, and 3D sensors. Fish factors (such as species and health) can be tracked, behavior can be analyzed, feeding choices can be made, and water quality can be predicted using this data. DL has recently emerged as a promising solution to the problems of ailing precision aquaculture production. To determine if salmon were eating, for example, the authors of [161] et al. used a dual-stream recurrent network; the findings demonstrated an accuracy rate of 80% for feeding categorization. However, many DL techniques need large, labeled datasets to train. Unfavorable conditions, including poor vision in muddy water, poor lighting, and cluttered backgrounds, may make it difficult to get high-quality data underwater. LLMs have emerged as a viable solution to help realize precision aquaculture farming because of their remarkable ability to handle diverse data from several domains utilizing few-short or zero-short learning. For example, generation-based LLMs may provide data and make decisions in fish aquaculture.

The lack of sufficient aquaculture data would exacerbate the difficulty of underwater species identification jobs. GANs have been used for image synthesis to improve visual identification in aquaculture, as depicted in Table 7. Two publicly available fish datasets, one from the Fish4-knowledge project (27,370 images of fish in 23 clusters) and the other from the Croatian fish dataset (794 images of 12 fish species), were used by Zhao et al. [162] to train a semi-supervised learning model for live fish recognition using a multi-class classifier in the discriminator. The semi-supervised GAN model outperformed both CNN-based models without data augmentation and other GANs (e.g., 48.39–59.44% for the standard DCGAN) in achieving fish identification accuracies of 80.52–83.07% using small fractions (5–15%) of labeled training samples.

4.1.2.4 Large language models in automated plant health analysis

Deep learning-based crop disease identification methods can automatically learn characteristics by constructing deep networks. Because of this, the semi-automatic issue with older machine vision technologies that need human involvement in feature extraction is now fully resolved. On the other hand, deep learning-based approaches are pretty dependent on dataset size. Additionally, the complicated application situations and high cost of data labeling make it challenging to gather significant amounts of high-quality labeled data in the agricultural field. A lack of data also hinders the development of agricultural automated identification. There have been substantial attempts to use generative adversarial networks (GANs) to supplement image data to

Table 7 Applications of large language models in precision agriculture

Reference	Application	Materials examined	Performance metrics (%)
Li and Tang [158]	Animal farming	Goats	Acc = 85.1
Singh et al. [159]		Cattle muzzles	Acc = 98.9
Ahmed et al. [160]		Poultry chickens	Acc = 97.0
Zhao et al. [161]	Aquaculture	Fish species	Acc = 80.5–83.1
Zhang et al. [162]		Shrimp eggs	Acc = 99.2

improve the identification of plant diseases and other health issues, thanks to their breakthroughs for image synthesis and their successful applications in disease diagnostics. The field of image augmentation for better plant disease detection has paid much attention to conditional generative adversarial networks (CGAN) and deep convolutional generative adversarial networks (DCGAN). In their study, Hu et al. [163] described DCGAN with conditional label constraint, called C-DCGAN, to detect red scab, red leaf spot, and leaf blight, three diseases that may affect tea leaves. Before input into classifiers (VGG16, SVM, decision tree, and random forest), original color images were enhanced with 20 to 4,990 additional images per illness using support vector machine (SVM) segmentation. Results showed that the visual geometry group (VGG16) achieved an average accuracy of 90% with the GAN-augmented data, which was 28% better than results obtained with rotation and translation-based augmentation and even better than results obtained with VGG16 devoid of data augmentation owing to extreme overfitting.

Abbas et al. [164] improved eleven tomato leaf image classes from the Plant Village dataset using CGAN. DenseNet121 achieved over 97% classification accuracy by adding 4000 synthetic photos per class to the training set. Compared to the original data without picture augmentation, these accuracies were around 1 to 4% greater. Douarre et al. [165] presented one of the first research to employ DCGAN to generate pictures for segmenting apple scab, a disease that manifests as spot lesions on the leaves and apples. After being split into 64 × 64-pixel sub-pictures, the infrared (IR) images acquired for the plant canopy were sent into SegNet. Compared to a baseline model without data augmentation, segmentation accuracy increased by 2% when DCGAN was used with the WGAN training approach. More significant progress was made using traditional data augmentation and model-based simulation approaches since DCGAN struggled to converge to produce realistic images. To identify a disease that caused the leaves to become yellow, Zeng et al. [166] collected 5406 color photos of citrus leaves from the Plant Village collection. By using DCGAN to create 8650 more images, the training samples were doubled, increasing the training data to 80% of the total. The classification accuracy increased by up to 20% using the GAN-enhanced data.

4.1.3 Few-shot learning of large language models in the agricultural domain

LLM contributes significantly to generating and processing agricultural data, offering insights and decision assistance. But when it comes to fully supervised learning, it often takes a lot of labeled examples to train to get a successful model. In contrast to many contemporary models, humans can learn new things even in the face of little or no experience. Few-shot learning (FSL) is a notion that academics have developed to bridge the gap between humans and huge models. To infer new information, FSL merely needs metaknowledge or past knowledge; in other words, it can also achieve impressive generalization capabilities based on sparse samples. It is very helpful in agriculture, where data collecting and labeling are costly and challenging. FSL, which goes by three different names: few-shot, one-shot, and zero-shot, has already shown exceptional performance in the NLP field [167]. Two to five samples per class are classified using a few-shot technique, one sample per class is classified using a one-shot approach, and invisible classes without samples are classified using a zero-shot approach. Tom B. Brown and colleagues trained and assessed GPT-3 in a few-shot scenario. According to the findings, GPT-3 excels in various NLP tasks, including cloze, translation, and quality assurance [168].

Unlike fully-supervised learning, FSL's generalizability is restricted, despite its strong performance in picture categorization, object recognition, and object tracking. Examine two instances of identifying and categorizing fruits in a picture. Although identifying fruits is a component of both jobs, they cannot be immediately transformed into one another. As a result, every task requires a specific model expressly trained in that particular goal, thereby diminishing the model's generalizability. Lack of labeled data is the primary issue with model training in the agricultural domain, which results in poor generalization of fully supervised learning-trained models. To lessen the need for specific crop data, researchers are attempting to employ FSL. Remarkably, LLMs have a natural aptitude for few-shot learning and do very well in generalizability [168]. LLMs have shown exceptional performance in various fields, including zero-shot generalization, without requiring task-specific fine-tuning. Nevertheless, these models are mostly restricted to handling textual data. MLLMs

may be able to get beyond these restrictions. It is important to note that many current research paths for big models ignore audio, video, and other components instead of concentrating just on graphics. GPT-4, OpenAI's most sophisticated big model, can only take text, audio, and picture input. Large, visually capable models demonstrate their adaptability in a variety of fields. These models are not effectively applied to the agricultural sector since the photos used to train them are taken from the Internet and differ significantly from actual agrarian photographs. Therefore, until there are appropriate agricultural pictures to train large vision-language models (LVLM), researchers studying agriculture can only depend on FSL to build a large model appropriate for agriculture.

Researchers use few-shot learning (FSL) in plant disease identification to train models to provide accurate predictions or classifications using a small quantity of labeled data or a support set [169]. Models in conventional machine learning often do better with massive data sets since this is how they discover patterns and correlations. Unfortunately, gathering and annotating giant data sets in real-world situations is not always feasible or cost-effective. Since convolutional neural networks (CNNs) rely on massive volumes of labeled data for efficient training, few-shot learning outperforms CNNs in situations with little data. Using few-shot learning may be very helpful when getting large labeled data sets, which is complex, such as when plant disease data sets are available. Models may provide good predictions with little data thanks to few-shot learning algorithms, which can adapt to new classes or categories with only a few samples. AI system prototype and development may be accelerated by few-shot learning. Without laborious data gathering and annotation, developers may quickly construct models for new tasks.

4.2 Large vision models in agricultural applications

Large vision models (LMV), or vision foundation models (VFMs), are LLMs with visual capabilities. LVM is a purely visual large model that trains and infers on picture data rather than language input [170]. These algorithms include key computer vision tasks such as classification, detection, segmentation, and generative modeling, and they have transformed standard computer vision approaches. These VFMs, trained on large picture datasets, may be easily converted to other domains, such as agriculture, with minimum fine-tuning and a small number of labels. Recently, the agricultural community has shown an increased interest in VFMs. LVM seeks to acquire universal visual knowledge and adapt to various visual activities and settings. The present use of big models in agriculture is mainly focused on CV. Using models to study diseases, pests, weeds, seeds, mature crops, and other elements requires using LVMs, with diseases and pests being the most pressing issue [171]. Traditional methods for diagnosing agricultural pests and illnesses depend heavily on unique techniques such as serology, molecular biology-based technology, and artificial visual assessment. Although these technologies may reliably identify pests and diseases to some degree, they are frequently time-consuming and costly. Furthermore, specific agricultural sample procedures often cause crop harm, contradicting the initial goal of identifying diseases and pests to safeguard crops. As a result, picture processing and analysis are critical tasks for big models in agriculture, such as embedding LVMs into robots to handle agricultural issues (weeding, branch trimming, harvesting, and so on) and accomplish automated agriculture.

4.2.1 Image processing and analysis

Using an LVM to judge crop-related information can significantly improve the time required for judgment and indirectly reduce the damage caused to crops. Moreover, after pests and diseases invade crops, their color, texture, and spectral characteristics will undergo specific changes related to computer vision. At present, there are four types of methods for obtaining crop image information: (1) Ordinary channels, taking photos to obtain images; (2) Obtaining remote sensing images through agricultural machinery near the ground; (3) Obtaining remote sensing images through aircraft monitoring platforms (4) Obtaining remote sensing images through satellites. Remote sensing can provide large-scale land use and land cover information. Various surface information can be identified by analyzing satellite or high-altitude images, such as surface conditions, soil moisture, vegetation coverage, and crop growth status [172].

Classifying and segmenting from limited examples obtained from remote sensing is a significant challenge. Wu et al. [173] proposed GenCo (a generator-based two-stage approach) for few-shot classification and segmentation on remote sensing and earth observation data. Their approach presents an alternative solution for addressing the labeling challenges encountered in remote sensing and agriculture domains. Spectral data can provide rich insights into the composition of observed objects and materials, especially in remote sensing applications. The challenges faced in processing spectral data in agriculture include: (1) Effectively processing and utilizing vast amounts of remote sensing spectral big data derived from various sources; (2) Deriving significant knowledge representations from intricate spatial-spectral mixed information; (3) Addressing the spectral degradation in the modeling of neighboring spectral relevance. Hong et al.'s SpectralGPT empowers intelligent processing of spectral remote sensing big data, and this LVM has also demonstrated its excellent spectral reconstruction capabilities in agriculture [174]. Multispectral imaging (MSI) and hyperspectral imaging (HSI) make monitoring crop health in the field possible. Integrating remotely sensed multisource data, such as HSI and LiDAR (Light detection

and ranging), enables monitoring changes occurring in different plant parts [175].

Using a large visual model to assess this spectrum data, crop health information may assist farmers in rapidly and effectively diagnosing and treating illnesses, lowering agricultural production loss. Using LVMs for picture identification and crop prediction is often more successful than machine learning techniques. When farmers require crop information, they may utilize one of four different picture-gathering techniques (Fig. 3). The image information is then processed through image recognition (which is divided into four tasks: image classification, object detection, semantic segmentation, and instance segmentation), and the identified results must be further predictively analyzed by the model (LVLM) to obtain crop information that farmers understand. In addition to gathering data by assessing crop phenotypic features, Feng et al. created an organelle segmentation network (OrgSegNet) [176]. OrgSegNet can correctly measure the diameters of chloroplasts, mitochondria, nuclei, and vacuoles inside plant cells, allowing for a more in-depth examination of plant phenotypes at the subcellular level. They've tried two apps: (1) A thermo-sensitive rice albino leaf mutant was grown in freezing temperatures. In transmission electron microscope (TEM) photos, the albinotic leaves lacked normal chloroplasts, and OrgSegNet failed to detect any chloroplast structures; (2) Young leaf chlorosis 1 (Ylc1). Young leaves of the ylc1 mutant had lower amounts of chlorophyll and lutein than the wild type, and TEM examination indicated a visible loose arrangement of thylakoid lamellar structures. Assume a huge model is utilized to replace deep learning methods. In such instances, subcellular cell detection may perform better, and the findings may be employed in subsequent predictive analytics to gather information that non-plant specialists can comprehend.

Yang et al. [177] used the segment anything model (SAM) to perform zero-shot chicken segmentation tasks utilizing both part-based and infrared thermal pictures. SAM beat other VFM algorithms, such as SegFormer and SETR, in whole and partial chicken segmentation. SAM obtained a mean Intersection over union (mIoU) of 94.80% in the chicken segmentation challenges. In contrast, SegFormer and SETR only produced mIoUs of 43.22 and 42.90%, demonstrating SAM's efficiency in poultry segmentation tasks. Furthermore, a novel single-bird tracking algorithm was proposed that combines SAM, YOLOX [178], and ByteTracker to track the activities of individual birds across video frames by extracting location information (i.e., bounding boxes provided by SAM) in each image and then tracking the birds over time with YOLOX and ByteTracker. Despite its promising results, the tracking algorithm lacked end-to-end functionality. The object tracking FMs [179] demonstrate outstanding performance in video object tracking and segmentation while needing little human intervention, making them potentially useful as an end-to-end technique. Williams et al. [180] presented an autonomous leaf segmentation pipeline known as "Leaf Only SAM," which was given for zero-shot segmentation of potato leaves. When compared to a fine-tuned Mask R-CNN model on the annotated potato leaf dataset, Leaf Only SAM had an average recall of 63.2% and an accuracy of 60.3%. These findings are lower than the Mask RCNN model, which has an average recall and accuracy of 78.7% and 74.7%, respectively. However, it is crucial to highlight that Leaf Only SAM does not need fine-tuning or an annotated dataset, demonstrating significant zero-shot generalization capabilities. These results imply that the direct deployment of FMs in a zero-shot approach may not always result in good performance owing to distribution changes.

4.2.2 Large-models in agricultural robot systems

A conventional agricultural robot system consists of perception, decision-making, and actuation modules compared to agricultural robots whose perception module utilizes computer vision and deep learning to accurately identify crops, soil conditions, and other relevant information. Agricultural robots, often referred to as "AgriRobots," are at the forefront of transforming modern agriculture. AgriRobots can automate various farming tasks such as planting, weeding, harvesting, and crop health monitoring. For instance, AI-guided robots can identify and remove weeds with precision, significantly reducing the use of herbicides. Similarly, drones with advanced imaging sensors can monitor fields for signs of disease or pests, enabling early intervention and reducing yield losses. The decision-making module utilizes this data to automatically provide suitable agricultural management strategies based on crop growth status and soil quality [181]. Nevertheless, traditional agricultural robot systems have limitations in processing large volumes of offline data. They lack high-performance data processing and high-quality actual-time control capabilities. This is due to the potential network communication and computing burdens associated with big data processing, causing decreased system performance and heightened costs [181]. Furthermore, they were usually designed for specific crops based on crop type and application requirements. The drawbacks of traditional systems were evident in their inflexible control logic and the absence of intelligence, including automatic decision-making and motion generation. In response to these challenges, it is necessary to use large models to help lift the intellectual features of agricultural robots. Table 8 below presents some valid advantages and disadvantages of traditional and large language robot systems.

However, like intelligent breeding and smart animal farming, the development and implementation of AgriRobots need a multidisciplinary approach that incorporates robotics,

Table 8 Advantages and disadvantages of traditional and large language model robot system

	Advantages	Disadvantages
Traditional agricultural robot systems	Relatively stable and reliable No need for large amounts of data Lower technical complexity Lower operating costs	No autonomous motion generation Lack of autonomous intelligence Limited adaptability
Generative intelligent agricultural robot systems	Autonomous motion generation Intelligent decision-making ability Strong adaptability	High technical complexity Requires a large amount of data support Higher operating costs

AI, plant science, environmental research, and data analytics. This combined strategy requires managing and analyzing considerable data from several areas, making it difficult [182]. However, the prospective deployment of LVMs may provide helpful guidance in addressing these difficulties. These models, trained on large-scale multimodal data, may offer the reliable prediction skills required for improved agricultural operations. For example, reinforcement learning big language models might allow for precise control of AgriRobots, while large language foundation models could give significant insights throughout the AgriRobot design process [183]. LVMs may provide an integrated solution by merging insights from several disciplines, creating more effective and efficient agricultural robots and transforming the future of agriculture.

Contemporary LVMs may be used in drones to assess crops and gather data on their growth, illness, yield, and other variables [184]. Moreover, ground machines equipped with LVMs may also facilitate the harvesting and classification of crops, in addition to enabling close-range insect detection. Yang et al. [185] proposed a latent variable model, the segment anything model (SAM), which utilizes infrared heat pictures for chicken segmentation tasks in a zero-shot manner. SAM may be used in agriculture to delineate immature fruits on a fruit tree and swiftly get yield estimation. Yang et al. [186] later introduced the track-anything model (TAM) by integrating SAM with video technology. Regrettably, TAM prioritizes the preservation of short-term memory over long-term memory. Nonetheless, TAM has significant promise within the agriculture sector. Enhancing its long-term memory capacity will enable it to detect early alterations in crop illnesses and provide timely alerts to farmers. Integrating LVMs like SAM and TAM into robotics may facilitate automation in agriculture, while these LVMs can also contribute to the automation of agricultural robot design.

Stella et al. [148] consulted ChatGPT and other LLMs to create the best robotic gripper for picking tomatoes. It should be noted that the ChatGPT version they used back then was GPT-3, which just allowed text input. The LVLM function is now available in ChatGPT, which has been upgraded to GPT-4. By using ChatGPT, designers of agricultural robots may partially automate the design process by inputting text descriptions and drawings. Traditional vision models are tiny, making them ill-suited to the real-time needs of farming robots, and the identification outcomes of agricultural pictures are often disappointing. Because pre-trained recognition results utilizing large-scale datasets frequently beat classic vision models, the introduction of LVM has overcome this impasse. The inference and prediction process usually takes more time and processing resources to finish, which leads to poor real-time performance. Nevertheless, LVMs' real-time performance may be somewhat enhanced by model architecture optimization, efficient inference methods, and hardware acceleration approaches. Prospective uses of LLMs in robotics research include improving human–robot interaction [187], task planning [187], navigation [188], and learning [188]. They can help robots learn and produce their natural language, which is excellent for following instructions, annotating data, and solving group problems. They allow robots to access and combine data from many sources, which may aid in continual learning. Because of this, robots may improve their performance in real time, learn new abilities, and adjust to new environments.

Despite obstacles like integration complexity and bias reduction, LLMs have begun to aid in testing environment simulations and provide opportunities for novel robotics research. Wu et al. [189] focus on personalizing robot household cleanup tasks. They demonstrate that robots can generalize user preferences from a small number of instances by integrating LLMs with language-based planning and perception. Users are asked to offer examples of item placement, which the LLM then uses to develop generalized preferences. To improve decision-making in real-world situations, Driess [190] presented an embodied approach that uses a transformer-based language model with embedded sensor inputs and language tokens to enable collaborative processing. Positive transfer from different training across language and visual domains is achieved when the model is trained end-to-end for distinct embodied tasks. LLMs have also been investigated as zero-shot human models to improve human–robot interaction further. The authors [191] et al. trained LLMs on large amounts of text data to outperform specialist machine-learning models in prediction performance for some HRI tasks. On the other hand, several limitations were noted, including a lack of robust reasoning

ability in spatial and numerical contexts and an oversensitivity to cues. By allowing LLMs to reason about sources of natural language input, the authors [192] create an "inner monologue" that improves LLMs' processing and planning abilities in robotic control situations. By integrating LLMs with different types of textual feedback, they can enhance the execution of user instructions in various domains, such as simulated and real-world robotic tasks involving mobile manipulation and tabletop rearrangement. This improves their decision-making process. Throughout these investigations, LLMs have served as the backbone for integrating common sense information into the operation of robotic systems.

Planning: Robotics relies more on LLMs for strategic planning [193]. Their natural language processing and generation skills improve human–robot interaction, allowing robots to comprehend and do complex tasks with spoken directions. LLMs are crucial to task planning, a higher-level cognitive process determining sequential actions to attain objectives. From autonomous production to home duties, comprehending and following multi-step instructions is essential.

Manipulation: LLMs improve robot manipulation [194], excelling in object detection, grasping, and teamwork. They assess visual and spatial information to identify the best way to interact with things, making them useful in precise, flexible operations like surgery and assembly line work. They improve real-world decision-making by integrating sensor inputs and verbal signals in an embodied framework. It helps the model learn and generalize from language and visual training data, improving its performance in embodied tasks.

Navigation: LLMs have transformed robotics navigation [195], enabling robots to explore complicated terrain accurately and quickly. LLMs excel in generating robot pathways and trajectories that account for complex environmental factors in motion planning. Accurate and dynamically adjustable navigation is essential in warehouses, transit, healthcare facilities, and intelligent homes. LLMs help localize and map, which are necessary for robot navigation. They allow robots to locate themselves in an environment while building or updating a spatial representation. Autonomous exploration, search and rescue, and mobile robot operations need spatial awareness. They have also improved collision-free navigation while accounting for obstacles and dynamic changes, making them crucial in situations where robots must follow predefined paths with accuracy and reliability, such as in automated guided vehicles (AGVs) and delivery robots.

4.3 Multimodal large language model and model assessment

4.3.1 Integration of multimodal models

Multimodal large language model (MLLM), or multimodal foundation models (MFMs), is a popular research topic that employs strong LLMs to solve multimodal challenges. LVLM supports graphics and text and is the most prevalent MLLM. Many academics have combined text, photos, audio, video, sensor data, depth information, point cloud [196], and more in recent years. Multimodal learning improves performance in agricultural applications [197, 198]. The farm community uses multimodal learning to optimize agricultural operations and increase results. Bender et al. published an open-source agricultural robotics multimodal dataset [199]. This dataset from cauliflower and broccoli farms supports agricultural robotics and machine learning research. Stereo color, thermal, hyperspectral, and meteorological and soil data are included. This massive dataset is crucial for farm robotics and machine learning innovation.

A unique cucumber disease identification method combining an MLLM with image-text-label information was proposed by Cao et al. [200]. They used multimodal picture text and image self-supervised contrastive learning to integrate label information from many domains. The method made sample distance measurements in image-text-label space easier. This novel method achieved 94.84% identification accuracy on a modestly extensive multimodal cucumber disease dataset. Bender et al. [199] created an open-sourced multimodal dataset from cauliflower and broccoli farms to support agricultural robotics and machine learning research. Stereo color, thermal, hyperspectral, and meteorological and soil data are included in the dataset. Cao et al. [197] suggested a multi-modal language model using image-text-label information for cucumber disease identification. They combined label information across domains with picture-text multimodal contrastive learning and image self-supervised contrastive learning. This approach measured sample distance in image-text-label space. Our innovative technique achieved 94.84% identification accuracy on a small multimodal cucumber illness dataset. MFMs have been considered in agriculture. Lu et al. [201] examined GPT 4, a multimodal language and vision model, for AGI. Robots used for fruit picking and agricultural monitoring might benefit from MFMs. However, present models generally use text-image data and are confined to QA activities. Few agricultural robotics applications use pictures, text, speech (human commands), and depth information (LiDAR or laser sensors). Multimodal data sources may improve the capabilities of agricultural robots used for fruit picking and crop monitoring [202]. Large models with low multimodality do fewer tasks and are less applicable.

4.3.2 Multimodal large language model in smart plant breeding

Smart or precision agriculture uses customized tactics to boost animal and agricultural yields and lower input.

costs. Data-driven agricultural operations use imaging, autonomy, and AI/robotics for crop health detection, weed identification and control, fruit detection, innovative crop management, intelligent plant breeding, and genetic progress. GANs (GANs are multimodal, large language models) have become more popular for refining machine-learning models using visual data in precision agriculture, as seen below. The groundbreaking method of smart plant breeding optimizes crop types using genotype, environment, and interaction data. It simulates plant features and performance (phenotype) using advanced "omics" technologies, AI, and big data [203]. This "smart breeding" promotes genetic gains, shortens breeding cycles, and creates plants adapted to changing conditions. Phenotype prediction accuracy depends on analyzing multidimensional or multimodal spatiotemporal omics data, including genotype, environment, and their interaction. In such a data-rich environment, typical machine-learning methods struggle. Multimodal foundation models (MFMs) may address this complexity. These algorithms can manage and learn from diverse data sources across different domains, improving plant phenotypic prediction accuracy and efficiency and opening up new opportunities in smart plant breeding. Multimodal foundation models merge visual, textual, and occasionally aural data to provide a holistic analytical lens, helping to understand complex agricultural issues, especially precision crop management.

4.3.3 Multimodal large language model in smart genetic advancement

An innovative method, "smart plant breeding," uses information about a plant's genetic makeup, surrounding environment, and the interplay between the two to produce better harvests. To forecast plant characteristics and performance, it builds successful models using advanced "omics" technology, artificial intelligence, and big data. This "smart breeding" aims to increase genetic diversity, decrease breeding times, and create plants that thrive in dynamic ecosystems. On the other hand, processing multidimensional or multimodal data, which includes genotype, environment, and their interaction, is essential for accurate phenotypic prediction and is derived from spatiotemporal omics. Traditional methods of machine learning face formidable obstacles in such a data-rich setting. Xu et al. [204] suggest that multimodal language models (MLMs) might solve this complexity. Developing these models, which can process and learn from many kinds of data in different domains, can revolutionize smart plant breeding by making more accurate and efficient phenotypic predictions.

4.4 Reinforcement learning large models (RLLMs) in agriculture

Deep reinforcement learning (DRL) is an emerging field that fuses deep learning with reinforcement learning to tackle decision-making tasks. Its impressive performance has been demonstrated across various applications, from computer games to intelligent transportation systems and robot control [205]. The agricultural community has recently shown heightened interest in DRL [206], leading to various innovative applications such as crop management, agriculture robots, and intelligent irrigation systems [207]. However, these methods often face generalization and sample efficiency challenges because they rely on learning tasks from scratch without comprehensive knowledge of vision, language, or other datasets. Although no published papers explicitly focus on applying reinforcement learning foundation models in agricultural scenarios, the field presents vast untapped opportunities for exploration and innovation.

4.5 General purpose applications of large language models

LLMs' versatility as general-purpose tools for untrained jobs is one of its most important uses. They naturally interpret, produce, and alter contextually appropriate human-like writing. This lets them do anything from language translation and question-answering to summarization, text production, and programming assistance [208]. LLMs may adjust to the style and tone of the text they analyze, making their outputs more user-friendly and context-aware. LLMs can be used as personal assistants to help users write emails or schedule appointments [209], in customer service to answer common questions to free up human resources for more complex issues, or to generate human-like text for websites based on prompts [210]. Data analysis relies on LLMs to filter massive amounts of text data, summarize essential points, and detect patterns people would take longer to find. Despite their many uses, LLMs, like any AI system, are only as good as their training data. We should utilize LLMs cautiously since they might inadvertently perpetuate biases in training data, resulting in unfair or erroneous findings.

4.6 Applications of large language models in medicine

Medical research and delivery are changing because of LLMs. LLMs are employed in clinical decision support

systems to provide doctors with evidence-based therapy recommendations [211, 212]. They may propose diagnoses, testing, and treatment plans by analyzing patient data and medical literature. LLMs may also be utilized in chatbots [213] to answer patient questions about symptoms or prescriptions, organize appointments, and give health advice. LLMs analyze much medical literature, uncover relevant studies, synthesize results, and anticipate future research trends [214]. LLMs may provide training materials, test questions, extensive explanations of complex medical concepts, and tailored feedback for medical students [215]. They may also imitate patient encounters to help trainees learn clinical skills. By analyzing media data, LLMs can help public health efforts predict illness outbreaks, track public attitudes to health policy, and effectively communicate health facts [216].

4.7 Applications of large language models in education

Incorporating LLMs into the educational sector presents chances to improve learning experiences, provide assistance for educators, and improve the production of instructional material. By assessing students' learning styles, performance, and preferences, LLMs may provide tailored study materials and practice questions to foster individualized learning experiences. LLMs may assist educators with developing lesson plans, grading assignments, and producing diverse and inclusive educational material, substantially increasing the time available for instruction and student engagement [217]. In language acquisition, LLMs function as sophisticated conversational partners, adept at modeling dialogues in several languages, rectifying grammatical errors, augmenting vocabulary, and assisting with pronunciation to facilitate fluency in practice [218]. Moreover, LLMs enhance educational accessibility by offering assistance to students with impairments. They can provide real-time transcriptions for those with hearing impairments, provide reading support for the visually impaired, and clarify complex materials for people with learning problems.

4.8 Applications of large language models in science

Like medical applications, LLMs may accelerate research by rapidly reviewing and summarizing scientific information. By providing clear and accessible research summaries, LLMs may aid researchers in remaining informed about the newest results, even in disciplines outside their competence. Furthermore, LLMs may assist scientists in generating novel ideas and study inquiries since their capacity to analyze extensive information enables them to reveal insights that may not be readily discernible to human researchers [219]. Furthermore, in scientific writing, LLMs may assist researchers in drafting publications, proposing enhancements, and ensuring compliance with specified formatting standards [219]. This not only conserves time but also enhances the clarity of scientific communication, facilitating more effective collaboration across multidisciplinary teams.

4.9 Applications of large language models in maths

Besides offering assistance for mathematical research and teaching, LLMs may aid in resolving mathematical difficulties by delivering step-by-step elucidations and directing users through intricate proofs and computations. They assist in identifying flaws in thinking or calculation and propose remedies, functioning as a vital resource for learning and verification. Large language models may be used to verify the correctness of mathematical proofs, serving as a first screening before human evaluation. Although they can not replace the diligent efforts of mathematicians, they may facilitate the process of proof verification [220]. Furthermore, LLMs improve accessibility to mathematics by converting intricate ideas and discoveries into comprehensible language for non-experts, thus bridging the divide between theoretical mathematics and practical applications in physics, engineering, and economics.

4.9.1 Applications of large language models in law

Large Language Models may facilitate the thematic analysis of legal texts by producing preliminary coding for datasets, detecting themes, and categorizing data based on these topics. This collaboration between legal specialists and LLMs has shown efficacy in studying legal materials, such as court decisions on theft, enhancing research efficiency and quality. Furthermore, LLMs have been assessed for their capacity to provide explanations of legal terminology, emphasizing the enhancement of factual correctness and relevance via integrating phrases from case law. Incorporating pertinent case law into the LLM enables the enhanced models to provide superior explanations with fewer factual inaccuracies [221, 222]. Furthermore, LLMs may be taught with particular topic expertise to execute legal reasoning duties and address legal inquiries.

4.9.2 Applications of large language models in finance

Large language models like BloombergGPT, trained on comprehensive private financial datasets, demonstrate enhanced efficacy in financial tasks. This underscores the need for domain-specific training in developing LLMs capable of comprehensively understanding and processing industry-specific terminology and ideas. The launch of FinGPT [223] as an open-source model provides clear and accessible resources for developing innovative applications, including

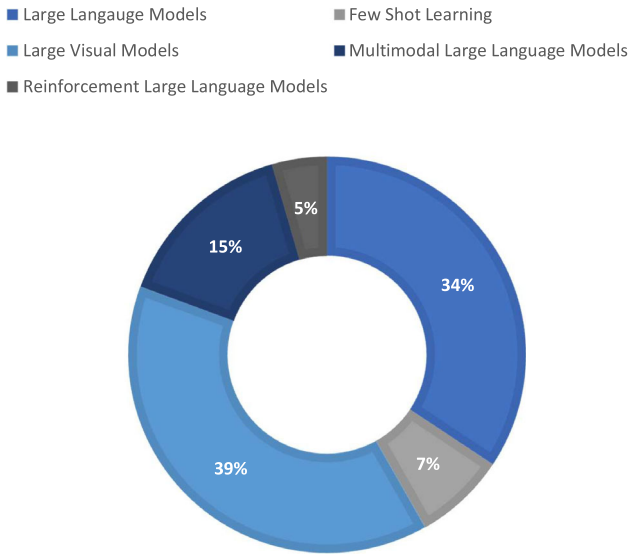

Fig. 13 Percentage uses of various large language models in different agriculture areas

robo-advising, algorithmic trading, and low-code solutions, enhancing financial services'.

capabilities. BloombergGPT and FinGPT demonstrate the versatility of large language models in the economic sector, with the former highlighting the efficacy of tailored datasets and the latter underscoring a data-centric methodology and low-rank adaption strategies for customization. Furthermore, LLMs may decompose complex financial problems into executable strategies, allowing comprehensive solutions [224].

4.9.3 Applications of large language models in coding

Various diverse applications of LLM models in coding are described in Sect. 2.3.2.5.

Figure 13 displays a statistical percentage representation of the various large language models in the smart and precise agricultural domain.

Similarly, Fig. 14 displays a statistical comparative representation of the applications of various large models in different application areas in addition to the agricultural domain.

Similarly, Fig. 15 displays a statistical representation of various large model types available in the recent literature.

5 Popular datasets for large language models

While large language models have shown encouraging results, the critical concern is how well they work and how to measure their effectiveness in particular activities or applications. The ever-changing nature of LLM applications makes their examination more difficult. Improving NLP activities like translation, summarization, question answering, etc., was the initial motivation for creating LLMs. Code generation and finance are only two examples of the many fields where these models already show usefulness. Fairness and prejudice, fact-checking, and logic are other important factors when evaluating LLMs. The present section describes the most popular metrics for evaluating LLMs. Different types of training and evaluation of the LLM capabilities are used to classify these benchmarks.

5.1 Datasets for language modeling, understanding, and generation

This section provides an overview of the benchmarks and datasets suited to evaluate the basic abilities of LLMs.

- Natural Questions [225] is one quality assurance dataset comprising actual, aggregated, anonymized queries typed into Google's search bar. An annotator finds a query and responds with the top most relevant pages. The page is labeled as null if neither of these things is present.
- MMLU [226] tests how well zero-shot and few-shot situations retain information, the benchmark for evaluating a model's general knowledge and problem-solving capacity. The primary use of this dataset is in multi-task thinking, question responding, and language comprehension.
- MBPP (Mostly Basic Python Problems) [227] is a standard for measuring how well code-generation models work. The benchmark covers a wide variety of subjects in its 974 brief Python programs, from basic programming ideas to standard library uses and beyond.
- HumanEval dataset [228] is used in tasks involving code creation, and it has 164 unique programming problems in its collection. The challenges come with their own set of documentation, code, and unit tests. Building this dataset to keep its contents out of code generation model training datasets was the primary motivation for its creation.
- Code-generating tasks focusing on Python are the specialty of APPS [229]. There are 232,444 Python applications in the APPS dataset, with an average of 18 lines of Python code in each application in the dataset. A library of 10,000 distinct programming exercises, each with a textual explanation of the topic, is also available via APPS.
- WikiSQL[230], designed for code generation tasks with 87,726 pairs of properly annotated SQL queries and matching natural language queries derived from Wikipedia tables. There are 17,284 examples in the test set, 9,145 instances in development, and 61,297 examples in training sets of SQL queries.
- TriviaQA [231] is QA-focused that has almost 650,000 question–answer-evidence triples. This dataset contains

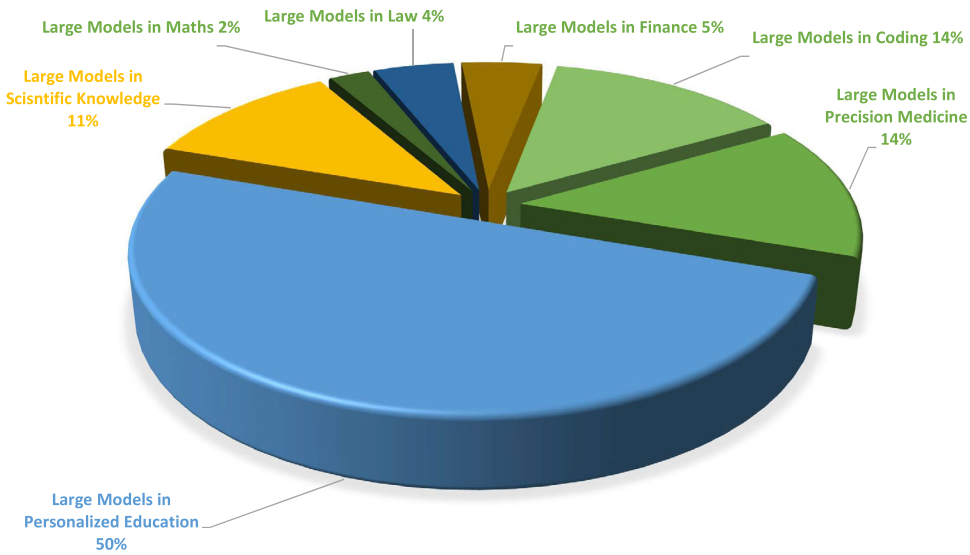

Fig. 14 Percentage comparison and uses of various large models in different application domains

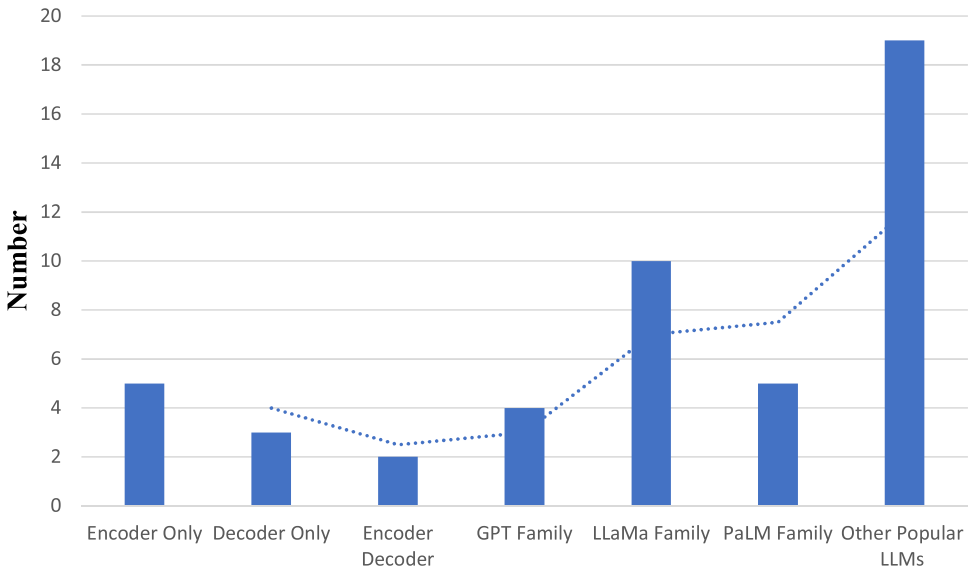

Fig. 15 Number representation of various large model types available in the recent literature. Data from Table 5

95,000 trivia enthusiast-authored question–answer pairings with an average of six independent proof documents.

- RACE [232] improves reading comprehension, comprising 28,000 texts and 100,000 questions meticulously produced by English teachers. Chinese middle and high school students aged 12–18 took English examinations in this dataset. This dataset includes several topics designed to test students' understanding and thinking. This dataset has three subgroups: RACE-M, RACE-H, and RACE. Middle school exams are RACE-M, whereas high school exams are RACE-H. Finally, RACE synthesizes RACE-M and RACE-H.
- The "Stanford Question Answering Dataset" [233] is a crowdsourced Wikipedia-based reading comprehension dataset with about 100,000 question–answer pairs linked to over 500 publications. These questions are usually answered with reading passage fragments or spans.
- Yes/no question-answering dataset BoolQ [234] aims to improve reading comprehension. BoolQ has 15,942 examples wherein, for example, triplets comprise a query, relevant text, and answer. This dataset is intended for reading comprehension but may also be used for reasoning, natural language inference, and question-answering.
- Another good reading comprehension dataset is MultiRC [235], which includes short paragraphs and multi-sentence questions to answer the paragraph. This dataset comprises passages from news, literature, historical writings, Wikipedia articles, society and law debates, elementary school science textbooks, and 9/11 reports. It comprises

about 6,000 multi-sentence questions from 800 paragraphs, each averaging two correct answers out of five [236].

- MATH [237] contains 12,500 high school math competition problems that are essential in evaluating the math skills of the model. Additionally, the standard score represents the severity of each problem, which is a number from 1 to 5 given to each problem. ′1′ symbolizes the simplest issues in a topic, whereas ′5′ represents the hardest.
- HellaSwag [238] tests the common sense of LLM models, which contains 70,000 multiple-choice questions. Each question comes from ActivityNet or WikiHow and offers four possible outcomes. The proper response describes the forthcoming occurrence, whereas the three erroneous answers mislead the machines.
- AI2 Reasoning Challenge (ARC) [239] is a benchmark dataset encompassing 7,787 scientific exams, most of which are English multiple-choice questions. Each collection is pre-divided into train, development, and test subsets, and a Challenge Set of 2,590 challenging questions and an Easy Set of 5,197 questions are included.
- The physical sense of the knowledge of the language representations' is assessed using PIQA [240]. This dataset prioritizes unusual answers for ordinary circumstances, employing multiple-choice questions (q) and two answers (s1 and s2) for every job. A model or person picks the best answer.
- SIQA [241] comprises 38,000 multiple-choice questions about daily emotional and social intelligence that has an advantage in assessing the social reasoning skills of the model. It uses a combination of human-selected and adversarially filtered machine-generated answers.
- OpenBookQA (OBQA) [242] is a novel question-answering dataset that demands deep text comprehension and common knowledge not found in the book. The questions were created using multi-stage crowdsourcing and expert filtering. These questions require multi-hop reasoning with little context, making them tough.
- TruthfulQA [243] evaluates language models' veracity in answering questions, and it contains 817 author-written questions from 38 areas, including health, law, money, and politics. These questions challenge humans since they may include frequent misconceptions that lead to erroneous replies.
- The diversified and explainable question-answering dataset in HotpotQA [244] requires multi-hop reasoning. This dataset comes from English Wikipedia and has 113,000 questions. Each dataset question has two gold paragraphs from two Wikipedia entries.
- ToolQA [245] is a question-answering benchmark that evaluates LLMs' ability to use external tools to answer questions.

6 Challenges of large language models

To fully capitalize on LLMs' potential in agriculture, it is necessary to have a thorough understanding of the various training and evaluation challenges and the existing opportunities. LLMs have proven to be highly effective in producing realistic images and enhancing deep learning models in these applications, but there are still a lot of potential challenges that need to be solved.

6.1 Challenges in smart agriculture

6.1.1 Data collection

Due to agriculture's diverse and unpredictable nature—impacted by crop varieties, development phases, soil conditions, weather patterns, farming techniques, etc.—the data-collecting process for training LLMs in agricultural applications poses particular obstacles. The complexity and diversity of the dataset are a consequence of these elements, which makes collecting and standardizing it a tough undertaking. The first and foremost concern should be the veracity and completeness of the data obtained and its capacity to accurately reflect the vast variations present in various agricultural settings and stages of development. Getting high-quality data may be laborious, costly, and time-consuming when it becomes urgent for ground truth labeling in supervised learning situations and the many development phases of plants or livestock. Generative models that generate high-fidelity images and label-efficient learning techniques that eliminate the need for lengthy labeling are promising approaches to address these difficulties. Second, data ownership and privacy issues arise since most farms are privately held. Concerns about privacy and possible economic exploitation make farmers wary about sharing data. To overcome these obstacles and provide more people with access to varied, high-quality agricultural datasets, new frameworks like federated learning (FL) have emerged. Another major obstacle is the dynamics of time. Every aspect of agriculture is subject to change, whether daily, seasonal, or yearly. This adds another degree of complexity as it requires gathering time-series data, such as data spanning years.

6.1.2 Distribution shift

One major obstacle to using LLMs in farming is the problem of distribution shift. A distribution shift occurs when there is a big difference between the data used to train the model and the data it encounters upon deployment. Regarding agriculture, the circumstances under which data is gathered might range substantially between regions and seasons. These changes

may significantly affect data distribution, including differences in crop kinds, soil conditions, weather patterns, and agricultural techniques.

6.1.3 Efficient training with limited data

The seemingly abundant training dataset powers LLMs' high-quality images. When training datasets are few, training LLMs to produce meaningful visuals for downstream deep-learning model construction may be difficult for agricultural applications like plant disease detection. Zhu et al. [246] created orchid seedlings using LLM models with varying training data counts. The images with the fewest training shots were visually poorer since they couldn't catch leaf and root features. Low-resolution images may lack texture details and have strange artifacts. After attempting to produce lemon images, Bird et al. [247] found that many synthetic images looked like potatoes, and a few had fake checkerboard patterns. Giving GANs additional data to train will help them create more realistic images, benefiting deep learning models. Insufficient data for LLM training leads the discriminator to overfit the training samples, making its input worthless to the generator and diverging the training process. Traditional image augmentation methods may help LLM training and reduce discriminator overfitting. In their work [248], authors employed LLM training to produce synthetic images for CNN model training, followed by data augmentation (flipping, rotation, and contrast adjustment). With this strategy, much research has achieved great image classification accuracy. Karras et al. [249] proposed an adaptive discriminator augmentation strategy that stabilizes training and improves LLM image formation with minimal training data. The authors also noted that the augmentation does not substitute high-quality LLM training data.

6.1.4 Large language models evaluation

LLM performance evaluation is complex. There aren't any common LLM performance or sample quality standards currently. The generated and real datasets do not match; thus, root mean square error (RMSE) does not function here. Qualitative visual assessments of generated images' perceptual quality are commonly used in computer vision literature, although they are subjective, vulnerable to human error, time-consuming for large datasets, and may overlook distributional aspects. Borji et al. [250] examined the most prominent metrics for comparing generated and actual images: Inception Score (IS), F1-score, precision and recall, Frechet Inception Distance (FID), mode score, and image structure similarity measures [251]. Other criteria might be used to assess visual quality and diversity. Few studies have used quantitative methods to evaluate LLM-generated images for agricultural applications and examined these measures' generalizability for agricultural imagery [164]. While it may be easier to evaluate LLMs on downstream machine learning tasks, future research should quantify the quality and variety of GAN-generated data to better understand their strengths and weaknesses and inspire LLM developments.

6.2 Ethical implications and responsible artificial intelligence

As Dark LLMs blur the limits between lawful and harmful usage, they create serious ethical considerations. Their exploitation endangers people, corporations, and society. Dark large language models (DLLMs) have complex ethical implications:

6.2.1 Algorithmic explainability and incorrect output errors

Since machine learning models base their decisions on probabilistic methods, Generative AI models tend to provide the fastest, most probable solution rather than the correct one. The outputs could mislead buyers as they seem more and more like the actual thing. When LLMs exhibit emergent behavior that is both irrational and erroneous while having plausible semantics or syntax, this is called a hallucination [252]. So, instead of using facts, the generative AI model uses its preconceptions and prejudices to produce content. Furthermore, verification of generative AI output is rare, especially LLM output. The efficacy of the learning process and the data used to train generative AI models are crucial. The use of accuracy checks in generative AI systems and applications can potentially prohibit certain outcomes. Customers must have trust in reliable findings since even cutting-edge AI algorithms are black boxes. The inability to modify and retrain models results from the closed-source nature of commercially available generative AI systems. Users may mitigate the potential downstream impacts of erroneous findings by verifying the explanations or references provided by generative AI. Despite their limitations, these probabilistic explanations may help consumers determine the reliability of generative AI outcomes.

6.2.2 Bias and fairness

Everyday, human-generated material is infused with societal prejudices. The alignment procedure and training data quality significantly affect vanilla generative AI's impartiality. By using biased data to train deep learning models, harmful language, gender, sexual orientation, political affiliation, and religious stereotypes may be reinforced. Multimodal generative AI models, like contrastive language-image pre-training (CLIP) [253] and the CLIP-filtered LAION dataset [254], are

fundamental to generative AI systems (e.g., Dall-E 2 or Stable Diffusion), although recent research has shown that these systems are biased. Human biases can seep into the models at different points during the model engineering process. The RLHF approach introduces more bias into instruction-based language models [255]. To mitigate these dangers, strict coding standards and quality checks are essential. Although there has been some progress in the scholarly literature on the topic, how to make AI fairer and freer of prejudice is still very much open and continuing. However, further research is required to come any closer to the idea of fair AI.

6.2.3 Blurred lines between legitimate and malicious use

It is difficult to differentiate between ethical and immoral uses of dark LLMs since they may be used for both good and evil. Because of this lack of clarity, worries over their abuse and unforeseen effects upon deployment have been voiced.

6.2.4 Risks to individuals, businesses, and society

Risks to people, companies, and society at large are associated with Dark LLM exploitation. Businesses risk intellectual property theft or privacy breaches, while individuals are susceptible to manipulation and disinformation. Dark LLMs can potentially damage society's faith in technology and AI systems.

6.2.5 Impact on fairness and equity

Unfair discrimination and socioeconomic inequities may worsen when dark LLMs, like other AI systems, reinforce biases in their training data. The results produced by Dark LLMs may reflect racial, gender, linguistic, and cultural biases, which might worsen existing societal inequalities.

6.2.6 Copyright violation

Due to its ability to generate results that are similar to or even identical to pre-existing works without acknowledgment to the sources, generative AI models, systems, and applications may lead to a breach of copyright regulations. Here, two possible dangers of infringement are shared. One argument against generative AI is that it might infringe on artists' reproduction rights by producing pirated versions of works. Examples of situations where this may occur include training a generative AI on copyrighted original material and having the AI create duplicates of that material. That copyright-free training material is essential for developing generative AI systems is a common assumption. The most important thing to remember is that copyright infringement may still occur even if the generative AI has never seen a copyrighted work before. This happens, for instance, when the AI makes a logo that looks a lot like Adidas's but has never seen the real thing. The other side is that generative AI has the potential to infringe on authors' transformative rights by creating derivative works. Consequently, the ratio of generative AI systems' originality to their creativity raises legal concerns. Following this line of thinking, legitimate concerns exist about the ownership of any creations (including patents) created by generative AI.

6.2.7 Misinformation

LLMs might transmit false information on occasion. Ensuring they don't stray from the truth is similar to being a thorough editor in a newspaper, always reviewing the facts.

6.2.8 Environmental impact

It is similar to supplying electricity to a small town while training these AI brains. They greatly impact the environment due to the amount of energy they use. It's like attempting to run a marathon while simultaneously planting trees: we want to keep these technologies progressing and benefit our world. Finally, developing and using generative AI systems is associated with significant environmental concerns. This is because these systems are usually built around large-scale neural networks, meaning they consume a lot of electricity during development and operation and negatively impact the environment. For example, it is estimated that the carbon emissions from training generative AI models like GPT-3 were 552 t CO2, the same as the yearly emissions of hundreds of families' worth of CO2.

6.2.9 Transparency

Figuring out how LLMs get to their answers might be as difficult as unraveling a mystery. Building trust and holding people accountable requires making these procedures more transparent, like a magician exposing some of their tricks to prove no magic is happening.

6.2.10 Safety

Last but not least, preventing these AI models from being misused is critical. Having such a powerful weapon may be like having the ability to construct a home or destroy it in an accident. Use it carefully so accidents don't happen. It is critical to put ethical concerns first in developing, deploying, and using LLMs because of the complexity of these technologies. Developing fair, open, and accountable AI techniques necessitates the establishment of regulatory frameworks and standards.

6.3 Security concerns

In cybercrime, LLMs have exceeded their primary objective of aiding human–computer interaction, which is scary. Cybercriminals using AI models to develop viruses, ransomware, and malware are a growing threat to internet safety. AI helps criminals write harmful code, uncover security weaknesses, and build up tempting traps for the unwary. Dark LLMs are emerging with unsettling powers and cybersecurity dangers. Evil large language models (LLMs) are called "Dark LLMs" and are utilized in cybercrime, fraud, cyberattacks, and deceptive content. Dark LLMs use powerful AI models for nefarious intents, advancing criminality. There are many types of Dark LLMs in literature.

6.3.1 XXXGPT

It is a malicious ChatGPT for cybercrime that supports botnets, remote access trojans (RATs), Crypters, and stealthy malware, providing a serious cybersecurity threat [256].

6.3.2 Wolf GPT

It uses Python and large harmful datasets to create unusual malware. Its strength is attacker anonymity, allowing sophisticated phishing and making cybersecurity detection difficult.

6.3.3 WormGPT

Based on the GPT-J model from 2021, WormGPT is an expert in cybercrime and malware creation. It can format code, chat memory, and infinite characters.

6.3.4 DarkBARD

Cybercrime is a fertile ground for DarkBARD, a nasty version of Google's BARD AI that manages multilingual communications and uses real-time data from the open web to generate disinformation and deepfakes. It is skilled at launching DDoS attacks and ransomware because of its varied content production capabilities and interaction with Google Lens. Cybercriminals now have cutting-edge tools to plan complex assaults and avoid detection thanks to these Dark LLMs, which represent the shadow side of AI and highlight the need for increased cybersecurity measures. The emergence of Dark Large Language Models (DLLMs) marks a significant evolution in cybersecurity. Cybercriminals may use dark LLMs to launch sophisticated assaults, such as social engineering, phishing, and the creation of false information and emails. Dark LLMs magnify cyber dangers because of how well they imitate human discourse. Threat identification and mitigation attempts may be hindered by their ability to create large volumes of misleading material rapidly. Dark LLMs are a persistent problem for cybersecurity experts because they are always changing and spreading. We must be proactive and use strong defenses, threat intelligence, and cross-sector cooperation to counter these advanced threats.

6.4 Training instability and other considerations

As indicated, unlike traditional image augmentation methods, LLM models may produce low-quality images. Numerous application studies suggest that LLM-based image augmentation increases DL model performance, although this may not always be true when dealing with the original data or more traditional image augmentation approaches. However, there is strong evidence that LLM with conventional augmentation regularly outperforms the approach alone or the original and LLM-augmented data. This suggests LLM and other image augmentation methods may work together to improve DL model performance. Further research is needed to compare GANs with traditional image improvement in farming. Semi-supervised GANs use a few tagged images and plenty of unlabeled data to achieve model performance comparable to supervised training. Kerdegari et al. [257] performed semi-supervised GAN on marijuana images using half the labeled training data. Semi-supervised GANs should be considered if labeling huge data is difficult. Makeup removal, which removes backdrops and occlusions, may help GAN image production. Kierdorf et al. [258] utilized Pix2Pix GAN to filter leaves from berry images that permitted images without occluded berries for accurate berry counting.

Although many regularization strategies have been suggested to address training instability in GANs, such as changes to the loss function, spectral normalization, gradient penalty, and model designs, the issue persists. The generator does not learn the distribution of training datasets effectively, which leads to mode collapse or dropping [259], and vanishing gradients, which cause convergence failures, are two typical training failures. GAN solves the min–max game using a gradient descent approach to reach the Nash equilibrium. When the discriminator outperforms the generator, the discriminator will always distinguish between actual and created samples, causing the generator's grain sizes to disappear (or approach zero), slowing down or halting the learning process. Mode collapse, in contrast, makes it seem like the generator can't provide any training data. Probably the most significant problem with GANs is this. At worst, the generator may generate the same set of examples (full mode collapse) or very few modes of training samples (partial mode collapse). In particular, mode collapse harms visual identification tasks, limiting GAN's capacity to produce different images. When trained on complicated datasets, even the most cutting-edge GANs (such as BigGAN and Trans GAN) experience instability and cannot be relied on to reliably generate images of sufficient quality. In addition, hyper-parameter

adjustments may have a significant impact on GAN performance. Therefore, many non-trivial training methods, such as hyperparameter tweaking and network structure engineering, are usually required to get GANs to perform.

7 Summary and conclusions

In the last few decades, there has been a surge in the digitization and automation of farming. Advances in artificial intelligence, such as LLMs and Gen AI, may hasten the automation of agricultural processes. Beginning with an introduction to early pre-trained language models (such as BERT), this article reviews three prominent LLM families (GPT, LLaMA, and PaLM) and several illustrative LLMs. It offers a thorough evaluation of the design components of LLMs, including architectures, datasets, and training processes, and helps to summarize essential results in the current research that are presented as summaries and discussions throughout the article. Overall, this research aimed to examine where large models stand in smart agricultural applications that promise to revolutionize agriculture by improving generalizability and efficiency while decreasing dependence on massive labeled datasets. Research using large language models in the agricultural sector can revolutionize the industry, paving the way for "smart agriculture" and "unmanned agriculture" while boosting production efficiency. Smart crop management, smart plant breeding, smart animal farming, precision aquaculture farming, and agricultural robots are some of the areas that may benefit from the many kinds of large pre-trained models that we have evaluated and classified.

We evaluate the results of many well-known pre-trained models using publicly available benchmarks and examine widely used LLM datasets. However, their use in smart agriculture remains in its infancy, as there are still a lot of unresolved issues in the development and deployment of large agricultural models that must be carefully considered before any models can be trained, validated, or deployed. While some studies have shown potential answers to these problems, further investigation is required to guarantee the usefulness and accuracy of large models in agricultural settings. Since the food industry and agriculture are so intertwined, developing large models for agriculture will inevitably lead to improvements in food models. When they interact, there will be positive feedback between these two domains of large models. For farmers' faith in new technology and the accuracy of the information they get, future research on large agricultural models should focus on making them more applicable and reliable. The potential for these models to improve agricultural decision-making, sustainability, and production warrants further investigation and use. This study aims to inspire and guide researchers to unlock the untapped potential of foundation models in smart agriculture. We hope this paper can serve as a support and cornerstone for developing future smart agricultural models with next-level intelligence and more sophisticated capabilities.

Funding This research has no funding source.

Data availability Data will be made public on request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval This article contains no studies with human participants or animals performed by any of the authors.

References

1. Li, J., Xu, M., Xiang, L., Chen, D., Zhuang, W., Yin, X., Li, Z.: Foundation models in smart agriculture: basics, opportunities, and challenges. Comput. Electron. Agric. **222**, 1–16 (2023)
2. Shaikh, T.A., Rasool, T., Verma, P.: Machine intelligence and medical cyber-physical system architectures for smart healthcare: taxonomy, challenges, opportunities, and possible solutions. Artif. Intell. Med. **146**, 102692 (2023)
3. Shaikh, T.A., Rasool, T., Lone, F.R.: Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. Comput. Electron. Agric. **19**, 107119 (2022)
4. Zhou, X., Ampatzidis, Y., Lee, W.S., Zhou, C., Agehara, S., Schueller, J.K.: Deep learning-based postharvest strawberry bruise detection under uv and incandescent light. Comput. Electron. Agric. **202**, 107389 (2022)
5. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y.: A survey on federated learning. Knowl. Based Syst. **216**, 106775 (2021)
6. Yang, J., Guo, X., Li, Y., Marinello, F., Ercisli, S., Zhang, Z.: A survey of few-shot learning in smart agriculture: developments, applications, and challenges. Plant Methods **18**(1), 1–12 (2022)
7. Li, J., Chen, D., Qi, X., Li, Z., Huang, Y., Morris, D., Tan, X.: Label-efficient learning in agriculture: a comprehensive review. Comput. Electron. Agric. **215**, 108412 (2023). https://doi.org/10.1016/j.compag.2023.108412
8. Goëau, H., Bonnet, P., Joly, A.: Overview of plantclef 2022: image-based plant identification at global scale. In CLEF 2022-Conference and Labs of the Evaluation Forum, 3180: 1916–1928. (2022)
9. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
10. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X. Gao, J.: Large language models: a survey, arXiv:2402.06196v2 [cs.CL], pp. 1–43, (2024)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. Dollár, P.: Segment anything, rXiv preprint arXiv:2304.02643, (2023)
12. Ahirwar, S., Swarnkar, R., Bhukya, S., Namwade, G.: Application of drone in agriculture. Int. J. Curr. Microbiol. Appl. Sci. **8**(01), 2500–2505 (2019)

13. Visentin, F., Cremasco, S., Sozzi, M., Signorini, L., Signorini, M., Marinello, F., Muradore, R.: A mixed-autonomous robotic platform for intra-row and inter-row weed removal for precision agriculture. Comput. Electron. Agric. **214**, 108270 (2023). https://doi.org/10.1016/j.compag.2023.108270
14. Abdullah, N.: Towards smart agriculture monitoring using fuzzy systems. IEEE Access **9**, 4097–4111 (2021)
15. Saleheen, M.M., Islam, M.S., Fahad, R., Belal, M.J., Khan, R.: IoT-Based smart agriculture monitoring system. In: Proceedings of IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, pp. 1–6, (2022), https://doi.org/10.1109/IICAIET55139.2022.9936826
16. Team, A.A., Bauer, J., Baumli, K., Baveja, S., Behbahani, F., Bhoopchand, A., Bradley-Schmieg, N., Chang, M., Clay, N., Collister, A. Dasagi, V.: Human timescale adaptation in an open-ended task space" arXiv preprint arXiv:2301.07608, (2023)
17. Geitmann, A., Bidhendi, A.J.: Plant blindness and diversity in AI language models. Trends Plant Sci. **28**, 1095–1097 (2023)
18. Kumar, S., Durai, S., Shamili, M.D.: Smart farming using machine learning and deep learning techniques. Decis. Anal. J. **3**, 100041 (2022)
19. Gzar, D.A., Mahmood, A.M., Adilee, M.K.A.: Recent trends of smart agricultural systems based on Internet of Things technology: a survey. Comput. Electr. Eng. **104**, 108453 (2022)
20. Vocaturo, E., Rani, G., Dhaka, V.S., Zumpano, E.: AI-driven agriculture: opportunities and challenges. In: 2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE, https://doi.org/10.1109/BigData59044.2023.10386314
21. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**(6), 1554–1563 (1966)
22. Katz, S.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. Acoust. Speech Signal Process. **35**(3), 400–401 (1987)
23. Mikolov, T., Karafiát, M., Burget, L.: Recurrent neural network based language model. Interspeech. **2**(3), 1045–1048 (2010)
24. Bengio, Y., Ducharme, R., Vincent, P.A.: Neural probabilistic language model. Adv. Neural. Inf. Process. Syst. **13**, 1–14 (2000)
25. Sundermeyer, M., Schlüter, R., Ney, H.: Lstm neural networks for language modelling. Interspeech. **2012**, 194–197 (2012)
26. Peters, M., Neumann, M., Iyyer, M. :Deep contextualized word representations ArXiv. (2018). https://doi.org/10.48550/arXiv.1802.05365
27. Vaswani, A., Shazeer, N., Parmar N.: Attention is all you need. Advances in neural information processing systems. (2017), 30.
28. Shanahan, M.: Talking about large language models. Commun. ACM **67**(2), 68–79 (2000)
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, vol. 25, pp. 1–25, (2012)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, vol. 25, pp. 1–17, (2012)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ArXiv. (2014). https://doi.org/10.48550/arXiv.1409.1556
32. Szegedy, C., Liu, W., Jia, Y.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on computer vision and Pattern Recognition, pp. 1–9, (2015)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 35: 770–778
34. Ren, S., He, K., Girshick, R.: Faster r-cnn: towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **28**, 1–18 (2015)
35. Redmon, J., Divvala, S., Girshick, R..: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, (2016)
36. He, K., Gkioxari, G., Dollár, P.: "Mask r-cnn". In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969, (2017)
37. Dosovitskiy, A., Beyer, L., Kolesnikov, A.: An image is worth 16 × 16 words: transformers for image recognition at scale" ArXiv. (2020). https://doi.org/10.48550/arXiv.2010.11929
38. Ramesh, A., Pavlov, M., Oh, G.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831, (2021)
39. Wu, J., Gan, W., Chen, Z.: Multimodal large language models: a survey. In: 2023 IEEE International Conference on Big Data (BigData), pp. 2247–2256, (2023)
40. Hoffmann, J., Borgeaud, S., Mensch, A.: Training compute-optimal large language models. ArXiv. (2022) https://doi.org/10.48550/arXiv.2203.15556
41. Le Scao, T., Fan, A., Akiki, C.: Bloom: a 176b-parameter open-access multilingual language model. ArXiv. (2023). https://doi.org/10.48550/arXiv.2211.05100
42. Anil, R., Dai, A., Firat, O.: Palm 2 technical report" ArXiv. (2023) https://doi.org/10.48550/arXiv.2305.10403
43. Zhang, S., Roller, S., Goyal, N.: Opt: open pre-trained transformer language models. ArXiv. (2022) https://doi.org/10.48550/arXiv.2205.01068
44. Zhu, D., Chen, J., Shen, X.: Minigpt-4: enhancing vision-language understanding with advanced large language models" ArXiv. (2023). https://doi.org/10.48550/arXiv.2304.10592
45. Zhao, L., Zhang, L., Wu, Z.: When brain-inspired ai meets agi. Meta-Radiology **1**(1), 100005 (2023)
46. Bubeck, S., Chandrasekaran, V., Eldan, R.: Sparks of artificial general intelligence: early experiments with gpt-4", ArXiv. (2023). https://doi.org/10.48550/arXiv.2303.12712
47. Gao, P., Han, J., Zhang, R.: Llama-adapter v2: parameter-efficient visual instruction model. ArXiv. (2023). https://doi.org/10.48550/arXiv.2304.15010.
48. Team, G., Anil, R., Borgeaud, S.: Gemini: a family of highly capable multimodal models. ArXiv. (2023). https://doi.org/10.48550/arXiv.2312.11805
49. Girdhar, R, El-Nouby, R.A., Liu, Z.: Imagebind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180–15190, (2023)
50. Wu, C., Lin, W., Zhang, X.: PMC-LLaMA: toward building open-source language models for medicine ArXiv. (2023) https://doi.org/10.48550/arXiv.2305.10415.
51. Driess, D., Xia, F., Sajjadi, M.S.M.: Palm-e: an embodied multimodal language model" ArXiv. (2023). https://doi.org/10.48550/arXiv.2303.03378.
52. Bai, J., Bai, S., Yang, S.: Qwen-vl: a frontier large vision-language model with versatile abilities ArXiv. (2023). https://doi.org/10.48550/arXiv.2308.12966.
53. Wu, S., Irsoy, O., Lu, S.: Bloomberggpt: a large language model for finance. ArXiv. (2023) https://doi.org/10.48550/arXiv.2303.17564.
54. Bi, Z., Zhang, N., Xue, Y.: Oceangpt: a large language model for ocean science tasks ArXiv. (2023) https://doi.org/10.48550/arXiv.2310.02031.

55. Wang, W., Dai, J., Chen, Z.: Internimage: exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14408–14419, (2023)
56. Liu, H., Li, C., Wu, Q.: Visual instruction tuning. Adv. Neural. Inf. Process. Syst. **36**, 1–17 (2024)
57. Dai, W., Li, J., Li, D.: Instructblip: towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems, vol. 36, pp. 1–121, (2024)
58. Wu, C., Yin, S., Qi, W.: Visual chatgpt: talking, drawing and editing with visual foundation models" ArXiv. (2023). https://doi.org/10.48550/arXiv.2303.04671
59. Ye, Q., Xu, H., Xu, G.: mplug-owl: modularization empowers large language models with multimodality. ArXiv. https://doi.org/10.48550/arXiv.2304.14178. 666, (2023)
60. Huang, S., Dong, L., Wang, W.: Language is not all you need: aligning perception with language models. Adv. Neural. Inf. Process. Syst. **36**, 1–11 (2024)
61. Gong, T., Lyu, C., Zhang, S.: Multimodal-gpt: a vision and language model for dialogue with humans" ArXiv. (2023) https://doi.org/10.48550/arXiv.2305.04790.
62. Wei, T., Zhao, L., Zhang, L.: Skywork: a more open bilingual foundation model" ArXiv. (2023) https://doi.org/10.48550/arXiv.2310.19341.
63. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205, (2023)
64. Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. Advances in neural information processing systems, vol. 13, (2000)
65. Schwenk, H., D´echelotte, D., Gauvain, J.-L.: Continuous space language models for statistical machine translation. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 723–730, (2006)
66. Mikolov, T., Deoras, A., Povey, D., Burget, L., Cernock, J.: Strategies for training large scale neural network language models. In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, pp. 196–201, (2011)
67. Cho, K., Van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder ap proaches," arXiv preprint arXiv:1409.1259, (2014)
68. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, (2018)
69. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. , Roberta: a robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, (2019)
70. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: a lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942, (2019)
71. Clark, K., Luong, M.-T., Le, Q. V., Manning, C. D.: Electra: pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, (2020)
72. Lample G., Conneau, A.: Cross-lingual language model pretraining," arXiv preprint arXiv:1901.07291, (2019)
73. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: generalized autoregressive pretraining for language understanding. Adv. Neural. Inf. Process. Syst. **32**, 1–29 (2019)
74. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.-W.: Unified language model pre-training for natural language understanding and generation. Adv. Neural. Inf. Process. Syst. **32**, 1–23 (2019)
75. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training, pp. 1–22, (2018)
76. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 1–19 (2019)
77. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
78. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: a massively multilingual pre-trained text-to-text transformer," arXiv preprint arXiv:2010.11934, (2020)
79. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y.: Mass: masked sequence to sequence pre-training for language generation," arXiv preprint arXiv:1905.02450, (2019)
80. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: denoising sequence-to sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461, (2019)
81. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A.: Language models are few-shot learners. Advances in neural information processing systems, vol. 33, pp. 1877–1901, (2020)
82. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, (2021)
83. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W.: Webgpt: browser assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332, (2021)
84. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A.: Training language models to follow instructions with human feedback. Adv. Neural. Inf. Process. Syst. **35**, 27730–27744 (2022)
85. OpenAI, "GPT-4 Technical Report," https://arxiv.org/pdf/2303.08774v3.pdf, (2023)
86. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi‘ ere, B., Goyal N., Hambro, E., Azhar, F.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971, (2023)
87. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S.: Llama 2: open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288, (2023)
88. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T. B.: Alpaca: a strong, replicable instruction following model, Stanford Center for Research on Foundation Models. https://crfm.stanford.edu/2023/03/13/alpaca.html, vol. 3 (6), pp. 1–7, (2023)
89. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: ef f icient finetuning of quantized llms, arXiv preprint arXiv:2305.14314, (2023)
90. Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., Song, D.: Koala: a dialogue model for academic research, Blog post, vol. 1, pp. 1–19, (2023)
91. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L.: "Mistral 7b," arXiv preprint arXiv:2310.06825, (2023)
92. Patil, S.G., Zhang, T., Wang, X., Gonzalez, J.E.: Gorilla: large language model connected with massive apis, (2023)
93. Pal, A., Karkhanis, D., Roberts, M., Dooley, S., Sundararajan, A., Naidu, S.: Giraffe: adventures in expanding context lengths in llms, arXiv preprint arXiv:2308.10882, (2023)
94. Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K., Wadden, D., MacMillan, K., Smith, N.A., Beltagy, I.: How far

can camels go? exploring the state of instruction tuning on open resources, arXiv preprint arXiv:2306.04751, (2023)
95. Mahan, D., Carlow, R., Castricato, L., Cooper, N., Laforte.: Available: "stable beluga models. [Online]. (https://huggingface.co/stabilityai/StableBeluga2)
96. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S.: Palm: scaling language modeling with pathways, arXiv preprint arXiv:2204.02311, (2022)
97. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S.: Scaling instruction f inetuned language models, arXiv preprint arXiv:2210.11416, (2022)
98. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z.: Palm 2 technical report, arXiv preprint arXiv:2305.10403, (2023)
99. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P.: Large language models encode clinical knowledge, arXiv preprint arXiv:2212.13138, (2022)
100. Zeng, W., Ren, X., Su, T., Wang, H., Liao, Y., Wang, Z., Jiang, X., Yang, Z., Wang, K., Zhang, X.: Pangu-α : large-scale autoregressive pretrained chinese language models with auto-parallel computation, arXiv preprint arXiv:2104.12369, (2021)
101. Zhang, Z., Gu, Y., Han, X., Chen, S., Xiao, C., Sun, Z., Yao, Y., Qi, F., Guan, J., Ke, P.: Cpm-2: large-scale, cost-effective pre-trained language models. AI Open **2**, 216–224 (2021)
102. Yuan, S., Zhao, H., Du, Z., Ding, M., Liu, X., Cen, Y., Zou, X., Yang, Z., Tang, J.: Wudaocorpora: a super large-scale chinese corpora for pre-training language models. AI Open **2**, 65–68 (2021)
103. Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y.: Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," arXiv preprint arXiv:2107.02137, (2021). 9, 22
104. Lieber, O., Sharir, O., Lenz, B., Shoham, Y.: Jurassic-1: technical details and evaluation, White Paper. AI21 Labs, vol. 1, pp. 1–32, (2021)
105. Kim, B., Kim, H., Lee, S.-W., Lee, G., Kwak, D., Jeon, D. H., Park, S., Kim, S., Kim, S., Seo D.: What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers, arXiv preprint arXiv:2109.04650, (2021)
106. Wu, S., Zhao, X., Yu, T., Zhang, R., Shen, C., Liu, H., Li, F., Zhu, H., Luo, J., Xu, L.: Yuan 1.0: large-scale pre-trained language model in zero-shot and few-shot learning, arXiv preprint arXiv:2110.04725, (2021)
107. Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S.: Scaling language models: methods, analysis and insights from training gopher, arXiv preprint arXiv:2112.11446, (2021)
108. Wang, S., Sun, Y., Xiang, Y., Wu, Z., Ding, S., Gong, W., Feng, S., Shang, J., Zhao, Y., Pang, C. and Liu, J.: Ernie 3.0 titan: exploring larger scale knowledge enhanced pre-training for language understanding and generation, arXiv preprint arXiv:2112.12731, (2021)
109. Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J.: Gpt-neox 20b: an open-source autoregressive language model, arXiv preprint arXiv:2204.06745, (2022)
110. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T.: Opt: open pre-trained transformer language models, arXiv preprint arXiv:2205.01068, (2022)
111. Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., Tow, J.: Bloom: a 176b parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100, (2022)
112. Banerjee, S., Dunn, P., Conard, S., Ng, R.: Large language modeling and classical AI methods for the future of healthcare. J. Med., Surg. Public Health **1**, 100026 (2023)
113. Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O.: Glam: efficient scaling of language models with mixture-of-experts, In: International Conference on Machine Learning. PMLR, pp: 5547–5569, (2022)
114. Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V.: Using deep speed and megatron to train megatron-turning nlg 530b, a large scale generative language model, arXiv preprint arXiv:2201.11990, (2022)
115. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D.L., Hendricks, L.A., Welbl, J., Clark, A.: Training compute-optimal large language models, arXiv preprint arXiv:2203.15556, (2022)
116. Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., Peris, C., Rawls, S., Rosenbaum, A., Rumshisky, A.: Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model, arXiv preprint arXiv:2208.01448, (2022)
117. Tay, Y., Dehghani, M., Tran, V.Q., Garcia, X., Wei, J., Wang, X., Chung, H.W., Shakeri, S., Bahri, D., Schuster, T., Zheng, H.S. Ul2: unifying language learning paradigms. In: The Eleventh International Conference on Learning Representations, (2022)
118. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X.: Glm-130b: An open bilingual pre-trained model, arXiv preprint arXiv:2210.02414, (2022)
119. Ren, X., Zhou, P., Meng, X., Huang, X., Wang, Y., Wang, W., Li, P., Zhang, X., Podolskiy, A., Arshinov, G.: Pangu-Towards trillion parameter language model with sparse heterogeneous computing, arXiv preprint arXiv:2303.10845, (2023)
120. Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., Xiong, C.: Codegen: an open large language model for code with multi-turn program synthesis, arXiv preprint arXiv:2203.13474, (2022)
121. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman.: Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374, (2021)
122. Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A.: Competition level code generation with alpha code. Science **378**(6624), 1092–1097 (2022)
123. Pang, R.Y., He, H.: Text generation by learning from demonstrations, arXiv preprint arXiv:2009.07839, (2020)
124. Wang, Y., Le, H., Gotmare, A.D., Bui, N.D., Li, J., Hoi, S.C.: Codet5+: open code large language models for code understanding and generation, arXiv preprint arXiv:2305.07922, (2023)
125. Li, R., Allal, L.B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J.: Starcoder: may the source be with you!" arXiv preprint arXiv:2305.06161, (2023)
126. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V. and Stojnic, R., Galactica: a large language model for science, arXiv preprint arXiv:2211.09085, (2022)
127. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y.: Lamda: language models for dialog applications, arXiv preprint arXiv:2201.08239, (2022)
128. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: Bloomberggpt: a large language model for finance. arXiv preprint arXiv:2303.17564, (2023)

129. Zhang, X., Yang, Q.: Xuanyuan 2.0: a large Chinese financial chat model with hundreds of billions parameters. arXiv preprint arXiv:2305.12002, (2023)
130. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652, (2021)
131. Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T.L., Raja, A., Dey, M.: Multi-task prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207, (2021)
132. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G.B., Lespiau, J.B., Damoc, B., Clark, A.: Improving language models by retrieving from trillions of tokens. In: International Conference on Machine Learning. PMLR, pp. 2206–2240, (2022)
133. Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L.: Improving alignment of dialogue agents via targeted human judgments. arXiv preprint arXiv:2209.14375, (2022)
134. Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T.: Solving quantitative reasoning problems with language models. Adv. Neural. Inf. Process. Syst. **35**, 3843–3857 (2022)
135. Tay, Y., Dehghani, M., Tran, V.Q., Garcia, X., Wei, J., Wang, X., Chung, H.W., Shakeri, S., Bahri, D., Schuster, T., Zheng, H.S., N. Houlsby., D. Metzler.: Unifying language learning paradigms. arXiv preprint arXiv:2205.05131, (2022)
136. Biderman, S., Schoelkopf, H., Anthony, Q.G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M.A., Purohit, S., Prashanth, U.S., Raff, E. and Skowron, A.: Pythia: a suite for analyzing large language models across training and scaling. In: International Conference on Machine Learning. PMLR, pp. 2397–2430, (2023)
137. Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H. and Awadallah, A.: Orca: progressive learning from complex explanation traces of gpt-4. arXiv preprint arXiv:2306.02707, (2023)
138. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B. and Liu, Q.: Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045, (2023)
139. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K. and Silver, D.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, (2023)
140. Song, H., Zhang, W.-N., Hu, J., Liu, T.: Generating persona consistent dialogues by exploiting natural language inference. Proc. AAAI Conf. Artif. Intell. **34**(05), 8878–8885 (2020)
141. Stella, F., Della Santina, C., Hughes, J.: How can llms transform the robotic design process? Nat. Mach. Intell. **5**(6), 1–4 (2023)
142. Niranjan, P.Y., Rajpurohit, V.S. and Malgi, R.: A survey on chatbot system for agriculture domain. In: 2019 1st International Conference on Advances in Information Technology (ICAIT), pp. 99–103, (2019)
143. Wolfram, S.: Alpha as the way to bring computational knowledge superpowers to chatgpt. Stephen Wolfram Writings RSS, Stephen Wolfram, LLC, vol. 9, pp. 1–14, (2023)
144. G. Lu, S. Li, and G. Mai, "Agi for agriculture" ArXiv. 2023. https://doi.org/10.48550/arXiv.2304.06136.
145. Peng, R., Liu, K., Yang, P.: Embedding-based retrieval with llm for effective agriculture information extracting. from https://doi.org/10.48550/arXiv.2308.03107. unstructured data. ArXiv. (2023)
146. Qi, C.R., Su, H., Mo, K. and Guibas, L.J.: Pointnet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660, (2017)
147. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, (2022)
148. Zhu, H., Qin, S., Su, M., Lin, C., Li, A. and Gao, J.: Harnessing large vision and language models in agriculture: a review, pp. 1–54, https://doi.org/10.48550/arXiv.2403.11858
149. Ramesh, A., Dhariwal, P., Nichol, A.: Hierarchical text-conditional image generation with clip latent. ArXiv. (2022), 1(2):3. https://doi.org/10.48550/arXiv.2204.06125
150. Shen, Y., Song, K., Tan, X.: Hugginggpt: solving ai tasks with chatgpt and its friends in hugging face. Adv. Neural. Inf. Process. Syst. **36**, 1–25 (2024)
151. Radford, A., Kim, J.W., Xu, T.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518, (2023)
152. Ren, Y., Ruan, Y., Tan, X.: Fastspeech: fast, robust and controllable text to speech. Adv. Neural. Inf. Process. Syst. **32**, 1–35 (2019)
153. Kundu, N., Rani, G., Dhaka, V.S., Gupta, K., Nayaka, S.C., Vocaturo, E., Zumpano, E.: Disease detection, severity prediction, and crop loss estimation in MaizeCrop using deep learning. Artif. Intell. Agric. **6**, 276–291 (2022)
154. Dhaka, V.S., Kundu, N., Rani, G., Zumpano, E., Vocaturo, E.: Role of internet of things and deep learning techniques in plant disease detection and classification: a focused review. Sensors **23**(18), 7877 (2023). https://doi.org/10.3390/s23187877
155. Farooq, M.S., Sohail, O.O., Abid, A., Rasheed, S.: A survey on the role of iot in agriculture for the implementation of smart livestock environment. IEEE Access **10**, 9483–9505 (2022)
156. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: segment anything meets videos. arXiv preprint arXiv:2304.11968, (2023)
157. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, (2021)
158. Li, H., Tang, J.: Dairy goat image generation based on improved-self-attention generative adversarial networks. IEEE Access **8**, 62448–62457 (2020)
159. Priyanka Singh, K., Devi, J., Varish, N.: Muzzle pattern based cattle identification using generative adversarial networks. In: Tiwari, A., Ahuja, K., Yadav, A., Bansal, J.C., Deep, K., Nagar, A.K. (eds.) Soft Computing for Problem Solving: Proceedings of SocProS 2020, pp. 13–23. Springer Singapore, Singapore (2021). https://doi.org/10.1007/978-981-16-2709-5_2
160. Ahmed, G., Malick, R.A.S., Akhunzada, A., Zahid, S., Sagriand, M.R., Gani, A.: An approach towards iot-based predictive service for early detection of diseases in poultry chickens. Sustainability **13**(23), 13396–14009 (2021)
161. Maløy, H., Aamodt, A., Misimi, E.: A spatiotemporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. Comput. Electron. Agric. **167**, 105087 (2019)
162. Zhao, J., Li, Y., Zhang, F., Zhu, S., Liu, Y., Lu, H., Ye, Z.: Semi-supervised learning- based live fish identification in aquaculture using modified deep convolutional generative adversarial networks. Trans. ASABE **61**(2), 699–710 (2018)
163. Gensheng, H., Haoyu, W., Zhang, Y., Wan, M.: A low shot learning method for tea leaf's disease identification. Comput. Electron. Agric. **163**, 104852 (2019). https://doi.org/10.1016/j.compag.2019.104852

164. Abbas, A., Jain, S., Gour, M., Vankudothu, S.: Tomato plant disease detection using transfer learning with CGAN synthetic images. Comput. Electron. Agric. **187**, 106279 (2021)
165. Douarre, C., Crispim-Junior, C.F., Gelibert, A., Tougne, L., Rousseau, D.: Novel data augmentation strategies to boost supervised segmentation of plant disease. Comput. Electron. Agric. **165**, 104967 (2019)
166. Zeng, M., Gao, H., Wan, L.: Few-shot grape leaf diseases classification based on generative adversarial network. J. Phys. Conf. Ser. **1883**, 012093 (2021)
167. Oniani, D., Chandrasekar, P., Sivarajkumar, S.: Few-Shot learning for clinical natural language processing using siamese neural networks: algorithm development and validation study. JMIR AI. **2**, e44293 (2023)
168. Brown, T., Mann, B., Ryder, N.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
169. Pan, J., Xia, L., Wu, Q., Guo, Y., Chen, Y., Tian, X.: Automatic strawberry leaf scorch severity estimation via faster R-CNN and few-shot learning. Ecol. Inf. **70**, 101706 (2022)
170. Bai, Y., Geng, X., Mangalam, K.: Sequential modeling enables scalable learning for large vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22861–22872, (2024)
171. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. Advances in neural information processing systems, vol. 25, (2012)
172. Khanal, S., Kc, K., Fulton, J.P.: Remote sensing in agriculture—accomplishments, limitations, and opportunities. Remote. Sens. **12**(22), 3783–3813 (2020)
173. Wu, J., Hovakimyan, N., Hobbs, J.: Genco: an auxiliary generator from contrastive learning for enhanced few-shot learning in remote sensing. ECAI **2023**, 2663–2671 (2023)
174. Hong, D., Zhang, B., Li, X.: SpectralGPT: Spectral remote sensing foundation model. ArXiv. 2023. https://doi.org/10.48550/arXiv.2311.07113.
175. Omia, E., Bae, H., Park, E.: Remote sensing in field crop monitoring: a comprehensive review of sensor systems, data analyses and recent advances. Remote. Sens. **15**(2), 354–379 (2023)
176. Feng, X., Yu, Z., Fang, H.: Plantorgan hunter: a deep learning-based framework for quantitative profiling plant subcellular morphology. https://doi.org/10.21203/rs.3.rs-1811819/v1, (2022)
177. Yang, X., Dai, H., Wu, Z., Bist, R., Subedi, S., Sun, J., Lu, G., Li, C., Liu, T., Chai, L.: Samfor poultry science. arXiv preprint arXiv:2305.10254, (2023c)
178. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: exceeding yolo series in 2021" arXiv preprint arXiv:2107.08430, (2021)
179. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968, (2023a)
180. Williams, D., Macfarlane, F., Britten, A.: Leaf only sam: a segment anything pipeline for zero-shot automated leaf segmentation. arXiv preprint arXiv:2305.09418, (2023a)
181. Yu, L., Liu, S., Wang, F.: Strategies for agricultural production management based on land, water and carbon footprints on the Qinghai-Tibet Plateau. J. Clean. Prod. **362**, 132563 (2023)
182. Fountas, S., Mylonas, N., Malounas, I., Rodias, E., Hellmann Santos, C., Pekkeriet, E.: Agricultural robotics for field operations. Sensors **20**(9), 2672–2699 (2020)
183. Team, A.A., Bauer, J., Baumli, K., Baveja, S., Behbahani, F., Bhoopchand, A., Bradley-Schmieg, N., Chang, M., Clay, N., Collister, A., Dasagi, V.: Human timescale adaptation in an open-ended task space. arXiv preprint arXiv:2301.07608, (2023)
184. Ganeshkumar, C., David, A., Sankar, J.G., Saginala, M.: Application of drone technology in agriculture: a predictive forecasting of pest and disease incidence. In: Applying Drone Technologies and Robotics for Agricultural Sustainability, pp. 50–81, (2023)
185. Yang, X., Dai, H., Wu, Z.: Sam for poultry science. ArXiv. (2023) https://doi.org/10.48550/arXiv.2305.10254
186. Yang, J., Gao, M., Li, Z.: Track anything: segment anything meets videos. ArXiv. https://doi.org/10.48550/arXiv.2304.11968, (2023a)
187. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: generating situated robot task plans using large language models. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 11523–11530, (2023)
188. Zhong, T., Wei, Y., Yang, L., Wu, Z., Liu, Z., Wei, X., Li, W., Yao, J., Ma, C., Li, X., Zhu, D.: Chatabl: abductive learning via natural language interaction with chatgpt. arXiv preprint arXiv:2304.11107, (2023)
189. Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S., Funkhouser, T.: Tidybot: personalized robot assistance with large language models. arXiv preprint arXiv:2305.05658, (2023)
190. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T. and Huang, W.: Palm-e: an embodied multimodal language model. arXiv preprint arXiv:2303.03378, (2023)
191. Zhang, B., Soh, H.: Large language models as zero-shot human models for human-robot interaction. arXiv preprint arXiv:2303.03548, (2023)
192. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y. and Sermanet, P.: Inner monologue: embodied reasoning through planning with language models. in 6th Annual Conference on Robot Learning, 2022. [Online]. Available: https://openreview.net/forum?id=3R3Pz5i0tye.
193. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: program generation for situated robot task planning using large language models. Auton. Robot. **47**(8), 999–1012 (2023)
194. Chalvatzaki, G., Younes, A., Nandha, D., Le, A.T., Ribeiro, L.F., Gurevych, I.: Learning to reason over scene graphs: a case study of finetuning gpt-2 into a robot language model for grounded task planning. Front. Robot. AI **10**, 1221739 (2023)
195. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 10608–10615, (2023)
196. Chen, C., Du, Y., Fang, Z.:Model composition for multimodal large language models. ArXiv. (2024) https://doi.org/10.48550/arXiv.2402.12750
197. Cao, Y., Chen, L., Yuan, Y., Sun, G.: Cucumber disease recognition with small samples using image-text-label-based multi-modal language model. Comput. Electron. Agric. **211**, 107993 (2023)
198. Dhakshayani, J., Surendiran, B.: M2f-net: A deep learning-based multi-modal classification with high-throughput phenotyping for identification of overabundance of fertilizers. Agriculture **13**(6), 1238–1271 (2023)
199. Bender, A., Whelan, B., Sukkarieh, S.: A high-resolution, multi-modal data set for agricultural robotics: a Ladybird's-eye view of Brassica. J. Field Robot. **37**(1), 73–96 (2020)
200. Cao, Y., Chen, L., Yuan, Y.: Cucumber disease recognition with small samples using image-text label-based multi-modal language model. Comput. Electron. Agric. **211**, 107993 (2023)
201. Lu, Y., Chen, D., Olaniyi, E., Huang, Y.: Generative adversarial networks (gans) for image augmentation in agriculture: a systematic review. Comput. Electron. Agric. **200**(107208), 1–25 (2022)
202. Tao, Y., Zhou, J.: Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. Comput. Electron. Agric. **142**, 388–396 (2017)

203. Gangwar, A., Dhaka, V.S., Rani, G., Khandelwal, S., Zumpano, E., Vocaturo, E.: Time and space efficient multi-model convolution vision transformer for tomato disease detection from leaf images with varied backgrounds. Comput. Mater. Contin. **79**(1), 117–142 (2024)
204. Xu, M., Yoon, S., Fuentes, A., Yang, J., Park, D.S.: Style-consistent image translation: a novel data augmentation paradigm to improve plant disease recognition. Front. Plant Sci. **12**(3361), 1–26 (2022)
205. Chen, D., Hajidavalloo, M.R., Li, Z., Chen, K., Wang, Y., Jiang, L., Wang, Y.: Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic. IEEE Trans. Intell. Transp. Syst. **24**(11), 11623–11638 (2023). https://doi.org/10.1109/TITS.2023.3285442
206. Gandhi, R.: Deep reinforcement learning for agriculture: principles and use cases. Data Sci. Agric. Nat. Resour. Manag. **2022**, 75–94 (2022)
207. Zhou, N.: Intelligent control of agricultural irrigation based on reinforcement learning," Journal of Physics: conference series. IOP Publishing, vol. 1601, pp. 1–11, (2020)
208. Hadi, M.U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S.: Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. TechRxiv, (2023)
209. Dong, X.L., Moon, S., Xu, Y.E., Malik, K., Yu, Z.: Towards next-generation intelligent assistants leveraging llm techniques. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 5792–5793, (2023)
210. Pandya, K., Holia, M.: Automating customer service using langchain: building custom open-source gpt chatbot for organizations. arXiv preprint arXiv:2310.05421, (2023)
211. Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., Succi, M.D.: Evaluating chatgpt as an adjunct for radiologic decision-making. medRxiv, pp. 1–20, (2023)
212. Benary, M., Wang, X.D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D.: Leveraging large language models for decision support in personalized oncology. JAMA Netw. Open **6**(11), e2343689–e2343689 (2023)
213. Montagna, S., Ferretti, S., Klopfenstein, L.C., Florio, A., Pengo, M.F.: Data decentralization of llm-based chatbot systems in chronic disease self-management. In: Proceedings of the 2023 ACM Conference on Information Technology for Social Good, pp. 205–212, (2023)
214. Pal, S., Bhattacharya, M., Lee, S.-S., Chakraborty, C.: A domain-specific next-generation large language model (llm) or chatgpt is required for biomedical engineering and research. Ann. Biomed. Eng. **2**(3), 451–454 (2023)
215. Abd-Alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P.M., Latifi, S., Aziz, S., Damseh, R., Alrazak, S.A., Sheikh, J.: Large language models in medical education: opportunities, challenges, and future directions. JMIR Med. Educ. **9**(1), e48291 (2023)
216. De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G.P., Ferragina, P., Tozzi, A.E., Rizzo, C.: Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. Front. Public Health (2023). https://doi.org/10.3389/fpubh.2023.1166120
217. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hullermeier, E.: Chatgpt for good? on opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
218. Young, J.C., Shishido, M.: Investigating openai's chatgpt potentials in generating chatbot's dialogue for English as a foreign language learning. Int. J. Adv. Comput. Sci. Appl. **14**(6), 1–28 (2023)
219. Altmäe, S., Sola-Leyva, A., Salumets, A.: Artificial intelligence in scientific writing: a friend or a foe?. Reproductive BioMedicine Online, (2023)
220. Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R.J., Anandkumar, A.: Leandojo: theorem proving with retrieval-augmented language models. arXiv preprint arXiv:2306.15626, (2023)
221. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z.: Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiol. **100017**, 1–21 (2023)
222. Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D.: Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. arXiv preprint arXiv:2308.11462, (2023)
223. Yang, H., Liu, X.Y., Wang, C.D.: Fingpt: open-source financial large language models. arXiv preprint arXiv:2306.06031, (2023)
224. Li, Y., Wang, S., Ding, H., Chen, H.: Large language models in finance: a survey. In: Proceedings of the Fourth ACM International Conference on AI in Finance, pp. 374–382, (2023)
225. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguist. **7**, 452–466 (2019)
226. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: Proceedings of 9th International Conference on Learning Representations (ICLR), Vienna, Austria, pp. 1–27, (2021)
227. Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., & Le, Q.: Program synthesis with large language models. arXiv preprint arXiv:2108.07732, (2021)
228. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.T., Choi, Y., Liang, P. & Zettlemoyer, L.: QuAC: Question answering in context, In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 2174–2184, (2018)
229. Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D. & Steinhardt, J.: Measuring coding challenge competence with apps. https://arxiv.org/abs/2105.09938, (2021)
230. Zhong, V., Xiong, C., Socher, R.: Seq2sql: generating structured queries from a natural language using reinforcement learning. arXiv preprint arXiv:1709.00103, (2017)
231. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1601–1611, (2017)
232. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: large-scale reading comprehension dataset from examinations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 785–794, (2017)
233. Rajpurkar, P., Zhang, J., Lopyrev K., Liang P.: SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp: 2383–2392, (2016)

234. Clark, C., Lee, K., Chang, M.W., Kwiatkowski, T., Collins, M., Toutanova, K.: Boolq: exploring the surprising difficulty of natural yes/no questions. CoRR, vol: abs/1905.10044, (2019)
235. Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S.,Roth, D.: Looking beyond the surface: a challenge set for reading comprehension over multiple sentences. In: Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), (2018)
236. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C.: Training verifiers to solve math word problems. CoRR, vol. abs/2110.14168, Available: https: //arxiv.org/abs/2110.14168, (2021)
237. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the MATH dataset. CoRR, vol. abs/2103.03874, (2021)
238. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: Hellaswag: can a machine really finish your sentence?" arXiv:1905.07830v1, (2019)
239. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try arc, the AI2 reasoning challenge. CoRR, vol. abs/1803.05457, 2018, Available: http://arxiv.org/abs/1803.05457.
240. Bisk, Y., Zellers, R., Gao, J., Choi, Y.: PIQA: reasoning about physical commonsense in natural language. CoRR, vol. abs/1911.11641, Available: http://arxiv.org/abs/1911.11641, (2019)
241. Sap, M., Rashkin, H., Chen, D., LeBras, R., Choi, Y.: Socialiqa: commonsense reasoning about social interactions. CoRR, vol. abs/1904.09728, Available: http://arxiv.org/abs/1904.09728, (2019)
242. Mihaylov, T., Clark, P., Khot, T., Sabharwal, A.: Can a suit of armor conduct electricity? A new dataset for open book question answering. CoRR, vol. abs/1809.02789, Available: http://arxiv.org/abs/1809.02789, (2018)
243. Lin, S., Hilton, J., Evans, O.: Truthfulqa: measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958, (2021)
244. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R. & Manning, C.D.: Hotpotqa: a dataset for diverse, explainable multi-hop question answering. CoRR, vol. abs/1809.09600, 2018. Available: http://arxiv.org/abs/1809.09600, (2018)
245. Zhuang, Y., Yu, Y., Wang, K., Sun, H., Zhang, C.: Toolqa: a dataset for llm question answering with external tools. arXiv preprint arXiv:2306.13304, (2023)
246. Zhu, F., He, M., Zheng, Z.: Data augmentation using improved cdcgan for plant vigor rating. Comput. Electron. Agric. **175**, 105603 (2020)
247. Bird, J.J., Barnes, C.M., Manso, L.J., Ekárt, A., Faria, D.R.: Fruit quality and defect image classification with conditional GAN data augmentation. Sci. Hortic. **293**(5), 1–11 (2022)
248. Bi, L., Hu, L.: Improving image-based plant disease classification with generative adversarial network under limited training set. Front. Plant Sci. **11**, 583438 (2020)
249. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proceedings of 34th Conference on Neural Information Processing Systems vol. 33, pp. 12104–12114, (2020)
250. Borji, A.: Pros and cons of GAN evaluation measures: new developments. Comput. Vis. Image Underst. **215**, 103329 (2022)
251. Xu, M., Yoon, S., Fuentes, A., Yang, J., Park, D.S.: Style-consistent image translation: a novel data augmentation paradigm to improve plant disease recognition. Front. Plant Sci. **12**, 773142–773142 (2022)
252. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)
253. Wolfe, R., Banaji, MR., Caliskan, A.: Evidence for hypodescent in visual semantic AI. Evidence for hypodescent in visual semantic AI. In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, pp. 1293–1304, (2022)
254. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv:2110.01963, (2021)
255. "OpenAI (2023b) How should AI systems behave, and who should decide?" https://openai.com/blog/how-should-ai-systems-behave [Last Accessed 11 June 2024].
256. "https://ensarseker1.medium.com/4-horsemen-of-the-apocalypse-wormgpt-fraudgpt-xxxgpt-wolfgpt-bonus-evilgpt-5944372575b8", [Last Accessed 15 September 2024].
257. Kerdegari, H., Razaak, M., Argyriou, V., Remagnino, P.: Semi-supervised GAN for classification of multispectral imagery acquired by UAVs. arXiv preprint arXiv: 1905.10920, (2019)
258. Kierdorf, J., Weber, I., Kicherer, A., Zabawa, L., Drees, L. & Roscher, R.: Behind the leaves—estimation of occluded grapevine berries with conditional generative adversarial networks. arXiv preprint arXiv:2105.10325, (2021)
259. Durall, R., Chatzimichailidis, A., Labus, P. and Keuper, J.: Combating mode collapse in GAN training: an empirical analysis using hessian eigenvalues. arXiv preprint arXiv: 2012.09673, (2020)