Attention is All Large Language Model Need

Yuxin Liu*

School of Software, Shanxi Agricultural University, 030600 Jinzhong, China

Abstract. With the advent of the Transformer, the attention mechanism has been applied to Large Language Model (LLM), evolving from initial single-modal large models to today's multi-modal large models. This has greatly propelled the development of Artificial Intelligence (AI) and ushered humans into the era of large models. Single-modal large models can be broadly categorized into three types based on their application domains: Text LLM for Natural Language Processing (NLP), Image LLM for Computer Vision (CV), and Audio LLM for speech interaction. Multi-modal large models, on the other hand, can leverage multiple data sources simultaneously to optimize the model. This article also introduces the training process of the GPT series. Large models have also had a significant impact on industry and society, bringing with them a number of unresolved problems. The purpose of this article is to assist researchers in comprehending the various forms of LLM, as well as its development, pre-training architecture, difficulties, and future objectives.

1 Introduction

The concept of Artificial Intelligence (AI) was first proposed in 1956, and the development of AI has gradually evolved from being based on small-scale expert knowledge to being based on machine learning. Until the birth of Convolutional Neural Network (CNN) and LeNet-5 in the 1980s, in-depth research in fields such as Natural Language Processing (NLP) and Computer Vision (CV) has been made possible by the evolution of machine learning methods from early shallow models to deep learning models [1]. In 2013, the NLP model Word2Vec was born, which proposed for the first time the word vector model that converts words into vectors, so that computers can better understand and process text data [2]. In 2014, the Generative Adversarial Network (GAN), hailed as one of the most powerful algorithm models of the 21st century, was born, marking a new stage of deep learning research into generative models [3]. In 2017, Google subversively proposed Transformer architecture, a neural network structure based on self-attention mechanism, which laid the foundation for large model pre-training algorithm architecture [4]. In 2018, OpenAI and Google released GPT-1 and BERT large models respectively, which means that pre-training large models has become the mainstream in the field of NLP [5,6]. GPT-3, the Large Language Model (LLM) at the time, was introduced by OpenAI in 2020 and had a model parameter size of 175 billion [7]. Subsequently, more strategies such as Reinforcement Learning from Human Feedback (RHLF), code pre-training, and instruction fine-tuning began to emerge and were used to

* Corresponding author: liuyuxin56@sxau.edu.cn

further improve reasoning ability and task generalization. In November 2022, ChatGPT, powered by GPT3.5, burst onto the scene and quickly became an internet sensation thanks to its realistic natural language interaction and multi-scenario content generation capabilities. The GPT-4, a newly released multimodal pre-training model that was released in March 2023, has the ability to comprehend multimodal and generate multitype content [8].

It must be admitted that large models, with their huge number of parameters and complex artificial neural network models, have become a key driving force for the continued development of AI. They signify the transition from the era of specialized AI research to the era of general AI, marking a monumental shift of great significance. Meanwhile, the emergence of large models has a significant impact on economic growth. By enabling more efficient and accurate automation, large models can help companies increase productivity, reduce costs, and expand into new markets. Large models can also help enterprises and government agencies make intelligent decisions, predict future trends based on historical and real-time data, and provide scientific decision-making basis. In addition, large models can also be used to enhance scientific research. By analyzing large datasets and identifying patterns and trends, these models can help researchers make new discoveries and advance human understanding of various phenomena. This could have profound implications for fields such as medicine, biology, and physics.

However, the emergence of large models has had a profound impact on society and industry, but also brought many open issues and challenges. The solution to effectively protecting user privacy and data security while ensuring model performance is urgent. Establishing a reasonable accountability mechanism to deal with the possible damage caused by AI systems is necessary. Optimizing the generalization ability and the ability to learn from few samples of the model is crucial. And achieving better multi-modal fusion and efficient use of data is also a pressing issue. These problems need to be solved urgently.

This article first introduces the Transformer, which is extremely important for LLM, and then divides it into single-modal large models and multi-modal large models based on the data types processed by LLM. The single-modality large model is refined into a text large model, an image large model, and an audio large model. Afterwards, the article comprehensively introduced different types of LLM and their development history, and deeply analyzed the pre-training architecture of LLM. Then, taking the GPT series as an example, the data sets and evaluation metrics used for training were introduced, and the characteristics and future challenges faced by the GPT series were analyzed and discussed. Finally, the impact of large models on society and industry, as well as the open issues and challenges they face in the future, were discussed separately. This article aims to provide researchers with a clear framework for understanding the current development status and future trends of LLM.

2 Large Language Model

2.1 Transformer

The Transformer, introduced by Vaswani in 2017, marks a pivotal advancement in AI. Unlike traditional CNN and Recurrent Neural Network (RNN), its core innovation lies in the Self-Attention mechanism, enabling parallel processing of input sequences and effective capture of long-range dependencies.

In the Transformer architecture, take the classic translation task as an example, the input texts, both source and target languages, are first converted into vector sequences through Embedding layers. Positional Encoding is then added to these vectors to provide information about the position of each word in the sequence, as the Self-Attention mechanism itself is

position-agnostic. Next, the English vectors, which represent the source language, are processed by a stack of n Encoder layers. Within each Encoder layer, they sequentially pass through Multi-Head Attention and Feed Forward networks. The model can provide a more complete understanding of the input context by simultaneously attending information from different positions in the sequence through the Multi-Head Attention mechanism. Similarly, the Chinese vectors, which represent the source language, are processed by a stack of n Decoder layers. However, the first attention mechanism in each Decoder layer is Masked Multi-Head Attention, which prevents the model from attending to future positions in the sequence. This is crucial for autoregressive decoding, where the model generates the output sequence one token at a time. After processing through the Decoder layers, the English vectors with added attention information and the Chinese vectors are combined and fed into a second set of Multi-Head Attention and Feed Forward networks to calculate the final output of the Decoder. This output is then passed through a Linear layer and a Softmax function to obtain the probabilities of each possible token in the target language, thereby generating the translation result. Throughout the Transformer's processing pipeline, Add (Residual Connection) and Norm (Layer Normalization) operations are performed after each Multi-Head Attention, Feed Forward, and Masked Multi-Head Attention step. These operations help stabilize the training process and improve the model's ability to generalize to unseen data [4].

With its efficient sequence processing and generation capabilities, Transformer has become the preferred choice for LLM. The introduction of the Transformer has significantly pushed the boundaries of what's possible in AI, paving the way for future innovations and advancements in the field.

2.2 Single-Modal Large Language Model

2.2.1 Text LLM

Each task in the NLP field requires a large amount of labeled data, and models cannot be reused. However, the CV field benefits from the ImageNet dataset, allowing the use of pre-trained models to fine-tune downstream tasks. Inspired by transformer, companies like OpenAI have begun to research pre-trained models in the NLP field.

GPT-n, a series of LLM developed by OpenAI, has now become synonymous with generative AI.

The first model in this series was GPT-1, where GPT stands for Generative Pre-Training. Due to the lack of labeled data, GPT-1 trained the model in an autoregressive manner using a language model. The model architecture employed a Decoder-only approach, which means it only used and improved upon the Decoder from the Transformer. GPT-1 introduced pre-trained model to NLP, but downstream tasks still required collecting some data for fine-tuning [5].

GPT-2 introduced prompts, descriptions of tasks, enabling the pre-trained model to tackle all downstream tasks. In terms of model architecture, unlike GPT-1, GPT-2 moved the layer normalization to the input of each block and added a layer normalization after the self-attention mechanism of the last sub-layer. Additionally, the initialization parameters of the residual layers decreased as the number of layers increased [9].

Subsequently, OpenAI found that although GPT-2 excelled in unsupervised learning, it was still comparable to models trained through supervision. Therefore, OpenAI promptly released GPT-3. GPT-3 continued the idea of GPT-2 by incorporating few-shot learning into prompts, providing the model with a few examples, and introducing sparse self-attention mechanisms. In terms of performance, few-shot learning significantly outperformed zero-

shot and one-shot learning, and GPT-3's few-shot capability surpassed other fine-tuned models [7].

GPT-4 is the first model in the OpenAI GPT series to possess multimodal capabilities, capable of receiving images and texts, and producing text and image outputs. GPT-4 utilizes RLHF to align the model with human consciousness. While this does not improve the model's performance, it makes the model's responses more aligned with human expectations [8].

Up to now, the latest model is GPT-o1, which is trained using the Chain Of Thought approach, causing it to spend more time contemplating before answering questions to ensure the provision of the most complete answer possible. This method of deep learning and contemplation enables it to give more accurate and comprehensive answers when faced with complex problems, and it demonstrates reasoning abilities that are comparable to or even surpass those of human doctors.

BERT is a pre-trained model similar to GPT, but it adopts a fundamentally different approach compared to GPT.

BERT utilizes the Encoder layer of transformers for semantic extraction, capturing long-distance dependencies. In other words, each token can see not only the tokens before it but also those after it. Therefore, BERT is better suited for semantic extraction and comprehensively outperforms GPT-1 on traditional NLP tasks. After pre-training, BERT can be fine-tuned for specific tasks to adapt to them [6].

BERT's pre-training methods encompass two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM task enhances the model's language understanding by training it to predict masked words. During training, the model is asked to predict a randomly selected portion of the words in the input sequence, which aids in the learning of vocabulary and grammar structures within the language. The NSP task trains the model's understanding of relationships between sentences by determining whether two sentences are adjacent. The model is instructed to take input from pairs of sentences and predict if the second sentence will follow the first, which assists it in understanding the relationships between sentences and the coherence of language [6].

2.2.2 Image LLM

Due to the emergence of Transformer, RNN have faded into the background in the NLP field. In the era of large models, traditional CNN in the CV field appears to struggle particularly when dealing with large-scale data or complex tasks. Consequently, Transformer have now also become a significant driving force in advancing the development of the CV field.

Vision Transformer (ViT) treats images as a series of patches and captures global information through the self-attention mechanism, thereby achieving powerful visual representation learning capabilities. In contrast to CNN, ViT learns global correlations between image features by using the Transformer's self-attention mechanism rather than convolutional procedures to extract picture data.

First, the input image is divided into multiple small, fixed-size blocks (e.g., 16x16 pixels), and each block is processed as part of the input sequence. These image blocks pass through a linear projection layer to obtain a linear embedding for each patch, forming embedded vectors known as Patch Embeddings. Then, a positional encoding is added to the embedded representation of each patch to provide positional information. After that, a special Class Token is appended at the beginning of the sequence for subsequent classification tasks. The sequence, including positional encodings and the class token, is then fed into the Transformer encoder, where features are extracted through multiple layers of self-attention mechanisms and feedforward neural networks. Each layer deepens the model's understanding of this information. Finally, a Multilayer Perceptron (MLP) is used for category prediction [10].

Visual Graph Neural Network (ViG) proposes a method for representing images as graph structures and utilizes Graph Neural Network (GNN) to extract graph-level features, thereby improving the model's performance in handling complex visual tasks. Similar to ViT, they both segment images into patches as the basic units for processing.

The ViG architecture includes Grapher and Feed Forward Network modules. Grapher processes graph info via graph convolution, capturing global/local features. Feed Forward Network transforms node features, enhancing diversity and reducing over-smoothing. Stacking modules create various ViG models. In ViG, an image is divided into several small patches. Each patch is converted into a feature vector through a simple mapping process, and these feature vectors are regarded as nodes within the graph structure. Subsequently, a graph structure is constructed by connecting nearest-neighbor nodes, where the connectivity between nodes reflects their spatial proximity in the image [11].

2.2.3 Audio LLM

Currently, the impact of transformers on LLM is not only evident in the fields of NLP and CV, but also demonstrates immense potential and wide application prospects in the realm of speech interaction. Among the core technologies in the field of speech interaction are Automatic Speech Recognition (ASR) and Text-To-Speech (TTS).

Whisper, released by OpenAI, is an ASR model using an encoder-decoder Transformer architecture for end-to-end speech recognition. The encoder converts audio signals into high-dimensional features via self-attention, capturing key info. The decoder predicts text based on encoder output, enhanced by special tokens for various tasks. Whisper achieves near-human robustness and accuracy in speech recognition [12].

Tacotron, developed by Google for TTS, uses an end-to-end Seq2Seq architecture to generate natural-sounding speech from text. It employs attention-based encoder-decoder, where the encoder converts text into feature vectors, and the decoder predicts speech spectrograms. The decoder focuses on key text info via attention. A post-processing network optimizes the spectrogram for realistic speech [13].

2.3 Multi-Modal Large Language Model

Driven by the advancements in transformers, single-modal large model has evolved into Multi-Modal Large Language Model (MM-LLM). MM-LLM integrates multi-modal information such as text, image, and audio, enabling cross-modal semantic understanding and generation, thereby further enhancing the level of intelligence of the models.

2.3.1 CLIP

Contrastive Language-Image Pre-Training (CLIP), a multimodal pre-trained neural network, was released by OpenAI in 2021.

CLIP's main components are an image encoder and a text encoder. The image encoder is responsible for converting images into feature vectors. It can be a CNN or a Transformer. These architectures are capable of capturing key features from images and converting them into vector forms that can be used for subsequent computations. The text encoder is accountable for turning text into feature vectors. It is typically a Transformer model that can handle long-distance dependencies and generate text vectors corresponding to the image vectors.

The core of CLIP is contrastive learning, a method for learning similarity metrics. The fundamental idea is to learn the similarity or difference between different sample pairs within the same dataset by comparing them. In CLIP, contrastive learning is employed to train the

model to learn the interrelationship between vision and language. Specifically, the CLIP model maps images and text into the same representation space and trains by contrasting the similarities and differences between different image-text pairs, thereby learning feature representations with good generalization ability. During the training process, the CLIP model receives a batch of image-text pairs as input and attempts to pull matched image and text vectors closer together in a common semantic space, while pushing unmatched vectors away. This learning approach enables CLIP to capture deep semantic connections between images and text, achieving cross-modal understanding [14].

2.3.2 LlaVA

Large Language and Vision Assistant (LLaVA) is a multimodal model trained to follow language and image instructions for real-world tasks. It uses CLIP for image encoding and Vicuna for text understanding, connecting them via a linear layer for joint image-text understanding and generation.

LLaVA is primarily composed of three parts: a visual encoder, a language model, and a fusion module. The visual encoder, akin to the function of human eyes, is responsible for analyzing and extracting important information from images. It typically employs technologies such as CNN or ViT, which excel at extracting rich feature information from images, including shapes, colors, and objects. The language model is responsible for understanding and generating natural language, similar to the language processing areas in the human brain. It uses models like GPT, which can comprehend complex sentence structures and contexts, and generate natural and fluent text. The fusion module combines image features from the visual encoder with textual information processed by the language model to form a holistic understanding. It acts like a translator, converting visual information into a format that can be integrated with textual information, or matching descriptions in text with visual content.

LLaVA accepts text, images, and videos with text descriptions. The visual encoder extracts image features, while the language model processes text. The fusion module combines these features for a comprehensive understanding, enabling the model to generate a natural language response [15].

2.3.3 Kosmos-1

The contemporary Kosmos-1, a multimodal LLM created by Microsoft, has the ability to analyze image content, address visual puzzles, recognize visual text, pass visual IQ tests, and grasp natural language commands.

The powerful multimodal capabilities of Kosmos-1 primarily rely on two important theoretical foundations: emergent abilities and chains of thought. Emergent abilities refer to the model's spontaneous acquisition of advanced and complex functions or capabilities during training, which were not explicitly specified during the training process. Chains of thought involve incorporating reasoning steps during model training to equip the model with logical reasoning capabilities.

Kosmos-1 integrates multimodal inputs using various encoding methods in a single sequence, marked by specific start and end symbols. It can insert any number of images. For text, it uses an embedding query table; for images, it employs CLIP's ViT-L encoder and DeepMind's Flamingo resampling. Kosmos-1 relies on Microsoft's MAGNETO network, enhanced by the Sub-LN structure, which excels in both text and vision tasks [16].

3 Experiment

3.1 Evaluation Index

Table 1. The relevant information about model parameters and datasets for the GPT series.

Model	Estimated Parameters	Dataset Name	Dataset Size	Data Source/Characteristics	Remarks
GPT-1	150 million	Books Corpus	Not Specified	Contains a vast amount of unpublished books with a wide range of topics and styles	Aids in learning diverse language structures and expressions
GPT-2	1.54 billion	Web Text	Approximately 40GB	Derived from approximately 8 million highly upvoted articles on Reddit	Features diversity, real-time relevance, and interactivity
GPT-3	175 billion	Various Internet Text Materials	Approximately 45TB	Encompasses a wide variety of text materials from the internet, with high diversity and breadth	Capable of understanding and generating text on various topics and in various styles
GPT-4	1.8 trillion	Multimodal Dataset (Unspecified)	Approximately 13 trillion Tokens	Includes text and code data, with instruction-tuned data from ScaleAI and internal sources	Supports text and code generation, trained through different numbers of epochs

As shown in Table 1, there were significant changes in the parameters of the GPT-1 to GPT-4 models. From GPT-1's 150 million parameters, to GPT-2's 1.542 billion parameters, to GPT-3's 175 billion parameters, the number of parameters in the model has shown an exponential growth. With GPT-4, this growth trend continues, and its parameter count has reached an astonishing 1.8 trillion, which is dozens of times that of GPT-3. This increase in the number of parameters not only means that the model can handle more complex and large data sets, but also represents a significant improvement in the model's ability to understand and generate natural language. As the number of parameters increases, the model can better capture the subtle differences and contextual information in language, resulting in more accurate, fluent, and natural text generation. Therefore, it can be speculated that with the continuous iteration and upgrading of GPT series models, their applications in language processing, text generation, dialogue systems and other fields will become more extensive and in-depth. The continuous improvement of these models will bring us a more intelligent and convenient natural language processing experience.

3.2 Datasets

The dataset used by GPT-1 is BooksCorpus, a collection that contains a vast amount of unpublished books with a wide range of topics and styles, which aids GPT-1 in learning diversified language structures and expressions [5].

The dataset utilized by GPT-2 is WebText, which is derived from approximately 8 million highly upvoted articles on Reddit, totaling approximately 40GB of data. The WebText dataset boasts diversity, real-time relevance, and interactivity, capturing the latest trends and expressions in online language [9].

The dataset used by GPT-3 originates from various textual materials on the internet, reaching approximately 45TB in size, and possesses a high degree of diversity and breadth, capable of covering a wide range of topics and styles. This large-scale data training enables GPT-3 to understand and generate text on various topics and in various styles [7].

The dataset used by GPT-4 contains approximately 13 trillion tokens, along with millions of lines of instruction-tuned data from ScaleAI and internal sources. GPT-4's dataset also exhibits a high degree of diversity and breadth, encompassing various textual and code data. It has undergone different numbers of epochs of training for text-based and code-based data to ensure that the model can better understand and generate both types of data [8].

3.3 Analysis

GPT-1 was the first to apply Transformer architecture to the NLP field, utilizing unsupervised training methods and introducing the paradigm of pre-training followed by fine-tuning [5].

With a significant increase in model size, GPT-2 demonstrated stronger contextual understanding capabilities and showed powerful few-shot learning abilities [9].

GPT-3 saw an exponential growth in parameter count, greatly enhancing the model's generative capabilities and generalization performance. It also excelled in solving multi-task problems and demonstrated robust zero-shot, few-shot, and many-shot learning abilities [7].

GPT-4, benefiting from its unprecedented model parameters, is more efficient and rapid in processing vast amounts of text and language tasks, with higher accuracy and customizability. Most importantly, GPT-4 has become a true multi-modal pre-trained large model [8].

3.4 Discussion

GPT-o1 is the latest model from OpenAI, which excels in reasoning and is capable of solving complex scientific, mathematical, and programming problems. Its deep learning and reasoning abilities enable it to spend time contemplating issues like humans do, trying out different strategies, and correcting itself when necessary.

GPT will continue to iterate and upgrade with the development of technology and hardware, and it boasts vast application scenarios and significant commercial value. As society progresses and AI becomes more widespread, people will pay more attention to it and have increasing demands. GPT faces not only many opportunities but also numerous challenges in the future. Firstly, there is the issue of whether the training data GPT relies on is legal. Secondly, its performance in many professional fields needs to be improved. Thirdly, there are ethical and regulatory concerns when using it. Lastly, there is the technical bottleneck mainly characterized by the black-box nature of neural networks.

4 Future of LLM

4.1 LLM's Influence

In recent years, increasingly intelligent LLM has been continuously launched onto the market, and it cannot be denied that people have entered a new era. This chapter discusses the impacts of LLM on society and industry.

4.1.1 Impact of LLM on Society

The LLM is exerting a profound impact on society. Firstly, through NLP technology, the LLM can quickly understand users' query intentions and provide them with accurate and comprehensive information. This has greatly changed the traditional way people obtain information, improving the efficiency and quality of information acquisition. Secondly, the LLM possesses powerful knowledge representation and reasoning capabilities, enabling it to integrate vast amounts of knowledge into a unified model and provide users with personalized knowledge services. This facilitates the dissemination and sharing of knowledge, stimulating innovative ideas. Thirdly, the popularization and application of the LLM have had a far-reaching impact on the employment structure. On the one hand, it has created new job opportunities, such as model trainers and data annotators; on the other hand, it has also replaced some traditional job roles, such as customer service representatives and data analysts. This requires people to continuously learn new skills and knowledge to adapt to changes in the job market. Lastly, the applications of the LLM in fields such as education, healthcare, and transportation help to improve social service levels and enhance people's quality of life.

4.1.2 Industrial Significance of LLM

The application of LLM in the industrial field is also gradually showing its great potential and value. First, the LLM can process and analyze a large amount of industrial data, extract valuable information and patterns from it, and provide support for decision-making in the production process. Through intelligent algorithms, it can optimize production processes, reduce resource waste, and improve production efficiency. Secondly, the application of LLM promotes innovation and technological upgrading in the industrial field. It can provide data support and intelligent design for new product development, speeding up the time to market for new products. Thirdly, LLM can analyze product data to identify potential quality issues and areas for improvement, thereby providing strong support for enhancing product quality. In addition, it can also carry out personalized product design and production based on customer needs and feedback, so as to improve customer satisfaction and loyalty. Finally, the applications of LLM in environmental protection and energy management also helps to promote the green development of industry. By monitoring and analyzing energy consumption and emission data in real time, LLM can help enterprises develop scientific energy conservation and emission reduction plans to reduce environmental pollution and energy consumption.

4.2 Open Issues and Challenges

With the rapid development of LLM, open Issues and challenges have also emerged accordingly. This chapter discusses the potential open Issues and challenges that LLM may encounter in the future.

4.2.1 Open Issues

First, personal privacy protection has become a major issue. How to effectively protect user privacy while ensuring model performance and avoiding data leakage and abuse is an urgent problem to be solved. Secondly, when an AI system malfunctions or causes damage, the issue of attribution of responsibility becomes complex. How to clarify the subject of responsibility and establish a reasonable accountability mechanism is a problem that needs to be solved at the legal level. Third, generalization and few-shot learning. LLM performs well when there is a large amount of data but struggles on tasks that require a small number of examples or specific domain knowledge. Finally, today's LLM tends to be multi-modal integration, and there is an increasing demand for multi-modal models that can understand and generate content including text, images, and other media types. In the process of integrating multiple modalities into a single model, data collection, training, and evaluation are all issues that need to be addressed.

4.2.2 Challenges

First, LLM is usually trained on a huge corpus of text data from Internet sources, which inevitably includes sensitive information, illegal information, gender and racial discrimination, and other cultural biases. Secondly, the training of the LLM requires the use of a large number of hardware devices and energy sources, which not only has a significant loss on hardware devices such as CPUs and GPUs, but also consumes non-renewable energy and pollutes the environment. Third, in order for the LLM to adapt to specific tasks, it needs to be fine-tuned, that is, the model is trained on a smaller data set, usually requiring manual annotators to label examples, which makes the process both time-consuming and expensive. Finally, the problem of the opaque and difficult-to-understand decision-making process of neural networks has not been solved, resulting in a significant reduction in the credibility of the model.

5 Conclusion

With the continuous advancement of neural network technology and the Transformer architecture, the performance of LLM has been rapidly improving, surpassing human intelligence in certain domains. LLM excel not only in traditional text data processing but also in multimodal data processing, such as image recognition and audio analysis, significantly broadening their problem-solving scope and enabling a wider range of complex application scenarios. However, the rapid development of LLM is accompanied by a series of new challenges, including ensuring healthy technological development, balancing human and AI capabilities, and mitigating potential risks, which have become focal points of current social attention and urgent research topics.

This article systematically introduces the basic principles of the Transformer architecture and the unique advantages of various LLM. Taking the GPT series models as a typical example, it deeply analyzes key aspects such as dataset selection criteria, fine-tuning of the training process, and scientific setting of evaluation indicators, providing readers with valuable practical experience and theoretical insights. Additionally, the article explores the profound impact of LLM on society and industry, as well as the complex issues such as technical bottlenecks and ethical dilemmas that may arise in their future development.

These discussions not only help to comprehensively understand the potential value and challenges of LLM but also point out the direction and path for promoting continuous progress in this field. The aim of this article is to assist researchers in better understanding the architecture and principles of LLM through in-depth explanations and comprehensive

analysis, stimulating their enthusiasm for exploration and innovation. Simultaneously, it calls for deep reflection and rational discussion among all sectors of society on the relationship between humans and AI, aiming to contribute wisdom and strength to building a more harmonious and sustainable AI society.

References

1. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86(11), 2278-2324 (1998)
2. T. Mikolov, Efficient estimation of word representations in vector space. arxiv preprint arxiv:1301.3781 3781 (2013)
3. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. Advances in neural information processing systems. 27 (2014)
4. A. Vaswani, Attention is all you need. Advances in Neural Information Processing Systems (2017)
5. A. Radford, Improving language understanding by generative pre-training. (2018)
6. J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv preprint arxiv:1810.04805 (2018)
7. T. B. Brown, Language models are few-shot learners. arxiv preprint arxiv:2005.14165 (2020)
8. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, Gpt-4 technical report. arxiv preprint arxiv:2303.08774 (2023)
9. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9 (2019)
10. A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale. arxiv preprint arxiv:2010.11929 (2020)
11. K. Han, Y. Wang, J. Guo, Y. Tang, E. Wu, Vision gnn: An image is worth graph of nodes. Advances in neural information processing systems. 35, 8291-8303 (2022)
12. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR (2023)
13. Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Z. Chen, S. Bengio, Q. V. Le. Tacotron: A fully end-to-end text-to-speech synthesis model. arxiv preprint arxiv:1703.10135, 164 (2017)
14. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR (2021)
15. H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning. Advances in neural information processing systems. 36 (2024)
16. S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu. Language is not all you need: Aligning perception with language models. Advances in Neural Information Processing Systems. 36, pp.72096-72109 (2023)