Survey Report, 7, Nov, 2024
Final Report for AI3605

Opportunities and Challenges of Large Language Models in Industry Applications

Yuan Gao

Abstract

In recent years, large language models(LLMs) have gained gained significant attention not only in academic, but also in industry. With the rising demand of the LLMs and the incredible potential interest of the AI-powered applications, it occurs massive opportunities also challenges. These LLMs, including GPT-4, Gemini, Qwen, and other advanced models, have demonstrated there abilities to automate tasks such as writing, coding, analysis and more. They are also transforming fields like healthcare, education, marketing by enabling prominent personalization and efficiency. However, there still exists several challenges performing as obstacles to the development of LLMs, including data privacy, ethics, cost and more. In this survey, we will explore and discuss about both opportunities and chanllenges of LLMs in industrial applications, providing insights into current research and future directions for addressing these obstacles.

Keywords

Keyword1 — Keyword2 — Keyword3

Contents

Introduction 1
1 Overview of Core-tech in LLMs 2
1.1 Objective of Language Modeling 2
Transformer Architecture ▪ Self-Attention Mechanism ▪ Multi-Head Attention ▪ Positional Encoding
1.2 Feed-Forward Neural Network (FFN) 3
1.3 Training Objective 3
1.4 Optimization 3
1.5 Scaling and Fine-Tuning 3
2 Industry Application Scenarios 3
2.1 Content Creation 3
2.2 Chatbot 3
2.3 Healcare 4
2.4 Education 4
2.5 Finace 4
3 Opportunities 4
3.1 Ehancing Efficiency 4
3.2 Improving Quality 4
3.3 Expanding Market Scale 4
3.4 Personalized Service 4
4 Challenges 4
4.1 Data Privacy 4
4.2 Data Resources 4
4.3 Ethics & Bias 4
4.4 Costs 4
4.5 Regulatory & Legal Risks 4
4.6 Technical Limitations 4
5 Conclusion 4
Acknowledgments 5
References 5

Introduction

The field of natural language processing (NLP) has changed a lot with the rise of large language models (LLMs). Coming from years of research in computational linguistics and deep learning, LLMs are based on early work like the use of neural networks for language modeling in the late 1990s and the transformer-based architectures introduced by Vaswani et al. [1]. These steps led to the creation of models like GPT [2], BERT [3], and, more recently, GPT-4, Gemini, and Qwen, which now surpass humans in many language tasks.

In the beginning, LLMs were praised in academic settings for advancing research in linguistics and machine learning. Their uses were mostly experimental, focusing on benchmarks and competitions like GLUE and SuperGLUE. But as models grew larger and AI-powered tools became more common, their use spread beyond academia. Now, industries like healthcare and marketing use LLMs to transform their work. These models help automate tasks like creating content, programming, and making decisions.

After the coming of GPT-3, the whole industry make sense that the era of LLMs are coming. From 2020 to now, industry release their LLMs like mushrooms after rain, few of them make a significant success including Claude, Gemini, Ernie, LLaMA. While others are still trying their best to make their

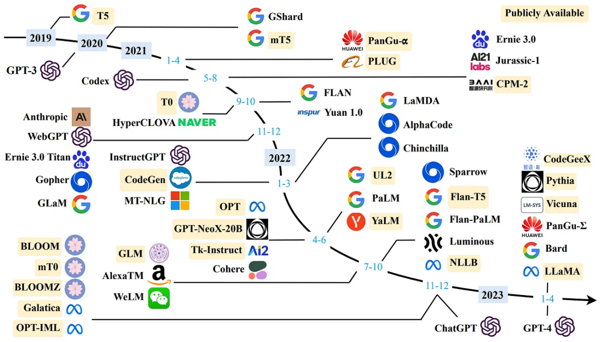

Figure 1. Chronological development of large language models (LLMs) from 2019 to 2023.

LLMs outstanding. We can briefly grasp the development context by Figure1

Even with their promise, using LLMs in industries comes with challenges. Problems like data privacy, ethical concerns, and the high cost of running these models often slow their wider use [4]. These issues not only limit their usefulness but also show gaps in research and implementation.

This survey aims to connect advances in research with real-world applications of LLMs. By collecting input from professionals and researchers, the study looks to find ways to use LLMs better in industries while addressing the problems that hold them back. The results aim to add to the discussion about AI's role in society and offer useful ideas for researchers, policymakers, and business leaders.

1. Overview of Core-tech in LLMs

1.1 Objective of Language Modeling

Large Language Models (LLMs) predict the probability of natural language sequences. Specifically, they compute the probability of the next word w_t based on previous words. This is expressed as:

$$P(w_t|w_1, w_2, \ldots, w_{t-1}).$$

For a full sequence $W = (w_1, w_2, \ldots, w_T)$, the model maximizes the likelihood function during training:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} \log P(w_t|w_1, w_2, \ldots, w_{t-1}; \theta).$$

Here, θ represents the model parameters.

1.1.1 Transformer Architecture

Transformer architecture is a neural network model. It has significantly impacted natural language processing (NLP). Unlike traditional recurrent neural networks (RNNs), transformers use an attention mechanism. This mechanism helps the model focus on important parts of the input sequence. This helps the model capture long-range dependencies. As a result, transformers perform better in various NLP tasks.

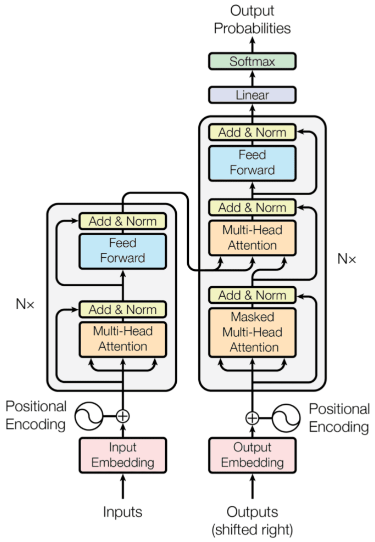

Figure 2. The encoder-decoder structure of the Transformer architecture Taken from “Attention Is All You Need“ [5]

1.1.2 Self-Attention Mechanism

The self-attention mechanism identifies the importance of input tokens relative to each other. For an input sequence $X \in \mathbb{R}^{n\times d}$, where n is the sequence length and d is the embedding dimension, the process follows these steps:

1. Compute query Q, key K, and value V matrices:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_k}$ are trainable weight matrices.

2. Calculate attention scores:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

The $\sqrt{d_k}$ term normalizes the dot product to improve stability.

1.1.3 Multi-Head Attention

Multi-head attention enhances the model's ability to learn diverse patterns. It divides self-attention into multiple parallel ”heads.” The output is:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W_O,$$

where each head is:

$$\text{head}_i = \text{Attention}(QW_{Q_i}, KW_{K_i}, VW_{V_i}).$$

Here, $W_{Q_i}, W_{K_i}, W_{V_i} \in \mathbb{R}^{d\times d_k}$ are weight matrices, and $W_O \in \mathbb{R}^{hd_k\times d}$ combines the outputs.

1.1.4 Positional Encoding

Transformers do not have recurrence. To handle token order, positional encoding is added to the input embeddings. With sinusoidal encoding, for position t and dimension i:

$$PE(t,2i) = \sin\left(\frac{t}{10000^{2i/d}}\right), \quad PE(t,2i+1) = \cos\left(\frac{t}{10000^{2i/d}}\right).$$

1.2 Feed-Forward Neural Network (FFN)

Each transformer layer includes a feed-forward neural network. It applies a non-linear transformation to each token independently:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2,$$

where W_1, W_2 are weight matrices, and b_1, b_2 are biases.

1.3 Training Objective

LLMs are trained using the negative log-likelihood of the true sequence:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P_\theta(w_t | w_1, w_2, \ldots, w_{t-1}).$$

The model computes $P_\theta(w_t|\cdot)$ using the softmax function:

$$P_\theta(w_t|\cdot) = \frac{\exp(z_t)}{\sum_{w' \in \mathcal{V}} \exp(z_{w'})}.$$

Here, z_t are the logits for token w_t, and $\mathcal{V}$ is the vocabulary.

1.4 Optimization

The model parameters are optimized using stochastic gradient descent (SGD) or its variants, like Adam. The gradient of the loss with respect to parameters θ is:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \log P_\theta(w_t | w_1, w_2, \ldots, w_{t-1})}{\partial \theta}.$$

The parameters are updated iteratively:

$$\theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta},$$

where η is the learning rate.

1.5 Scaling and Fine-Tuning

Fine-tuning is a technique that we use to adapt a pre-trained LLM to a specific task or domain. This allow us to train pre-trained LLM on a smaller, field-specific dataset, which can make the LLMs behave well in the specific domain. By fine-tuning, we can improve the model's performance, and more inspiringly, we can make experts in any specific field.

LLM performance improves with scaling. Key scaling strategies include:

- **Depth**: Increase the number of transformer layers.
- **Width**: Use larger embedding dimensions.
- **Data**: Train on large-scale text corpora.

By the escalation of the factors we mentioned above, LLM can grow stronger and stronger. Figuratively, LLM is just like a child who is learning to speak. With more knowledge(data), bigger brain(width), the child will become more and more smart. This is so called **Scaling Laws**, which is one of the most important factors that cause the rapid development and potential issues of LLM.

2. Industry Application Scenarios

With the development of LLM, there has been a various fields that embed LLM into their own applications. Here we will introduce some of the most common application scenarios.

2.1 Content Creation

Script Writing Film script writing is a creative process that combines art and skills, which traditionally requires a deep background and practical accumulation. Today, artificial intelligence and big data technologies have effectively improved its efficiency and quality. Recently, The Shanghai Jiao Tong University team proposed using the Large Language Model (LLM) to realize interactive drama [6], a new art form that combines traditional drama with modern AI technology. By redefining the six elements of drama, the audience can interact with the characters, explore and influence the development of the plot, and gain a richer experience.

2.2 Chatbot

Customer Support Using LLMs' powerful language capabilities, it's efficient to integrate specific knowledge, speaking manner, and other required information into chatbots. Therefore, chatbots can perfectly play the role of customer service staff. For example, Wofeng Technology provides Bayer (China)'s virtual medical representative platform with intelligent customer service products supported by AI big models, completing the AI empowerment of intelligent virtual representatives in the enterprise WeChat channel, thereby creating an intelligent customer service system for the expert community of the Academy of Imaging, helping Bayer (China) achieve a doubles model growth rate far higher than the offline representative singles and the industry average, academic refined operations with high recognition from doctors, access to nearly 100 hospitals throughout the year, and the formation of a discipline circle to continuously promote other products.

Q&A Systems When someone first encounter the LLM, if you ask them how to use it, the intuition will tell that Q&A systems are born for LLM. Figure 3 shows almost all of the current Q&A systems that are based on LLMs. The flourish of Q&A systems tells everything.

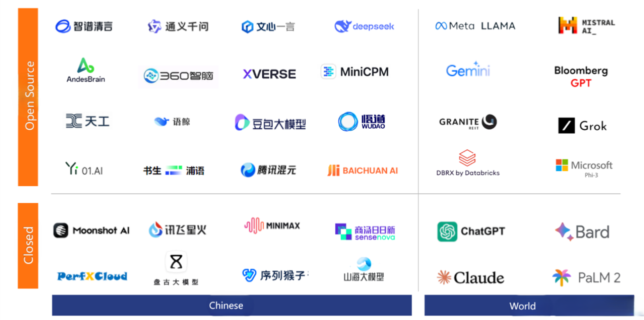

Figure 3. Various LLM applications around world

2.3 Healcare

Diagnostic Assistance On September 19, 2023, Baidu released the Chinese first “industrial-grade” large medical model - the ”Lingyi(Great Doctor)” big model based on LLM. In terms of diagnosis assistance, Baidu Lingyi Big Model provides doctors with efficient and accurate tools through functions such as medical record generation, patient condition briefing, intelligent question and answer, and clinical decision support. For example, it can automatically generate standardized medical records, extract key diagnosis and treatment information, answer professional questions based on authoritative literature, and connect with hospital systems to optimize workflows, thereby greatly improving medical service efficiency and diagnosis quality.

Pharmaceutical Huawei Cloud Pangu Drug Molecular Model has performed well in tasks such as molecule generation, property prediction and optimization by learning massive amounts of drug molecular data, greatly improving the efficiency of new drug research and development, shortening the research and development cycle and reducing costs. For example, the model helped develop broad-spectrum antibacterial drugs, shortening the research and development cycle from several years to one month and reducing costs by 70%, providing strong support for the intelligent transformation of the pharmaceutical industry.

2.4 Education

Personalized Learning In 2023, Chinese education technology companies actively applied big models in the field of education and launched a number of innovative applications to improve teaching and learning effects through intelligent means. In July, NetEase Youdao released the big model ”Zi Yue” for K12 education, which accomplishes personalized analysis guidance, guided learning and other functions. The big model can better teach students in accordance with their aptitude and provide students with all-round knowledge support. In August, TAL Education Technology released their big model MathGPT in the field of mathematics, which can automatically generate questions and give answers, covering elementary school to high school mathematics knowledge.

LLMs in the field of education are becoming a new tool for intelligent assisted teaching. Their knowledge integration capabilities can meet the dynamic needs of students, realize personalized learning, and improve the quality of teaching together with teachers.

Language Learning Named Large Language Models, LLMs are born to be brilliant in language field, which can be proved by the fact that the first batch of embedding LLMs into industrial applications is professional language teaching institutions like Duolingo. As early as 2021, before ChatGPT formally published, Duolingo has embeded GPT-3 into their language-teaching application to help generate learning content. As LLMs growing more and more powerful, now it can be used in many scenarios including DIY learning content, making personalized learning plans, helping students to improve their writing skills, and thanks to MLLMs(multi-modal LLMs), it can even be used to talk with learners and help them to improve their listening and oral skills.

2.5 Finace

Financial Analysis With wide range of financial data accumulated over past 40 years, Bloomberg released BloombergGPT, a LLM specifically designed for financial analysis. Because of the combination of powerful base model and high-quality finacial data, BloombergGPT behaves incredible in financial analysis tasks just like an experienced experts. This applications help to reduce the workload of financial analysts and improve the efficiency.

3. Opportunities

3.1 Ehancing Efficiency

3.2 Improving Quality

3.3 Expanding Market Scale

3.4 Personalized Service

4. Challenges

example: IBM Watson’s failure

4.1 Data Privacy

4.2 Data Resources

4.3 Ethics & Bias

4.4 Costs

4.5 Regulatory & Legal Risks

4.6 Technical Limitations

In this part, we can introduce RAG as an optional method to enhance LLM but still some issues

5. Conclusion

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio.

Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

Nulla mattis luctus nulla. Duis commodo velit at leo. Aliquam vulputate magna et leo. Nam vestibulum ullamcorper leo. Vestibulum condimentum rutrum mauris. Donec id mauris. Morbi molestie justo et pede. Vivamus eget turpis sed nisl cursus tempor. Curabitur mollis sapien condimentum nunc. In wisi nisl, malesuada at, dignissim sit amet, lobortis in, odio. Aenean consequat arcu a ante. Pellentesque porta elit sit amet orci. Etiam at turpis nec elit ultricies imperdiet. Nulla facilisi. In hac habitasse platea dictumst. Suspendisse viverra aliquam risus. Nullam pede justo, molestie nonummy, scelerisque eu, facilisis vel, arcu.

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam congue neque id dolor.

Donec et nisl at wisi luctus bibendum. Nam interdum tellus ac libero. Sed sem justo, laoreet vitae, fringilla at, adipiscing ut, nibh. Maecenas non sem quis tortor eleifend fermentum. Etiam id tortor ac mauris porta vulputate. Integer porta neque vitae massa. Maecenas tempus libero a libero posuere dictum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aenean quis mauris sed elit commodo placerat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Vivamus rhoncus tincidunt libero. Etiam elementum pretium justo. Vivamus est. Morbi a tellus eget pede tristique commodo. Nulla nisl. Vestibulum sed nisl eu sapien cursus rutrum.

Nulla non mauris vitae wisi posuere convallis. Sed eu nulla nec eros scelerisque pharetra. Nullam varius. Etiam dignissim elementum metus. Vestibulum faucibus, metus sit amet mattis rhoncus, sapien dui laoreet odio, nec ultricies nibh augue a enim. Fusce in ligula. Quisque at magna et nulla commodo consequat. Proin accumsan imperdiet sem. Nunc porta. Donec feugiat mi at justo. Phasellus facilisis ipsum quis ante. In ac elit eget ipsum pharetra faucibus. Maecenas viverra nulla in massa.

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetuer quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Acknowledgments

So long and thanks for all the fish [7, 8].

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *OpenAI Research*, 2018.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the NAACL-HLT*, 1:4171–4186, 2018.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[6] Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Jiale Hong, Hai Zhao, and Min Zhang. From role-play to drama-interaction: An llm solution, 2024.

[7] A. J. Figueredo and P. S. A. Wolf. Assortative pairing and life history strategy - a cross-cultural study. *Human Nature*, 20:317–330, 2009.

[8] J. M. Smith and A. B. Jones. *Book Title*. Publisher, 7th edition, 2012.