Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/35/20230399

The evolution, applications, and future prospects of large language models: An in-depth overview

Jiayin Li

School of Computing and Data Science, Xiamen University Malaysia, Negeri Selangor, 43900, Malaysia

IBU2209396@xmu.edu.my

Abstract. The evolution of natural language processing has transpired through three primary phases, with large-scale language models significantly transforming the field. These models have heightened the machine's capability to understand, produce, and interact with human language in unprecedented ways. Progressing from RNNs to transformer models, transitioning from encoder-decoder frameworks to decoder-centric designs, and the journey from BERT to the Chat-GPT series have marked significant shifts in the academic discourse. Impressively, these sophisticated models have infiltrated a range of sectors, including finance, healthcare, biology, and education, revolutionizing both traditional and emerging domains. However, as these advancements are celebrated, the ethical and economic challenges they introduce must also be addressed. Confronting these pivotal issues and harnessing technology for societal betterment has become a priority for academia and industry alike, sparking intense research endeavors in recent times. This review dives into the history of natural language processing, highlighting the pivotal developments and core principles of large language models. It provides a comprehensive perspective on their adoption and influence within the financial sector, crafting a detailed narrative of their deployment. In conclusion, the analysis reflects on the current challenges posed by these models and presents potential solutions. This study stands as a definitive guide, offering readers an in-depth understanding of the development, application, and future trajectories of large-scale language models.

Keywords: large language models, natural language processing, challenges and opportunities.

1. Introduction

A century ago, mathematician Andrei Markov found inspiration in the poetic prose of "Eugene Onegin," penned by Alexander Pushkin. This encounter led to the inception of the "language modeling" realm. Through analyzing sequences of words and phrases in the text, Markov discerned patterns, predicting subsequent words or phrases, and essentially mimicking Pushkin's linguistic style using probabilistic methods. This technique came to be known as the Markov Chain [1, 2], and its influence reverberated across diverse scientific and technological frontiers.

In the modern era, large-scale language models stand at the forefront of Natural Language Processing (NLP), reshaping the dynamics of our interaction with digital systems. Bolstered by advanced artificial intelligence methodologies, these models showcase unparalleled prowess in interpreting and reproducing human language nuances.

DOI: 10.54254/2755-2721/35/20230399

The ChatGPT series, in particular, heralds a transformative phase in NLP. It manifests an adeptness in orchestrating nuanced, context-aware conversations, a feat made possible due to rigorous training on expansive datasets. With its profound grasp of grammar, context, and semantic nuances, ChatGPT finds resonance across multifarious sectors. This detailed research exposition offers a panoramic view of the progression, utility, and prospective avenues of large-scale language models. It meticulously traces their evolutionary arc, elucidating the foundational technological tenets. This study also encapsulates the salient application spheres of these models, casting a spotlight especially on the financial sector, elucidating the myriad opportunities they present. In addition, the research delves into the inherent challenges that contemporary large language models present, offering potential mitigations from five distinct vantage points—intending to navigate these intricacies responsibly and sustainably. This exploration, encompassing technological intricacies, application spectra, and forward-looking insights, equips readers to fathom the monumental role this technology is poised to play in steering future human-digital dialogues. As NLP continues its metamorphic journey, this document stands as an indispensable compass for scholars, professionals, and aficionados venturing into the expansive realm of large-scale language models.

2. Research on large language model technology

The advent of large-scale language models represents a revolutionary leap forward in Natural Language Processing (NLP), empowering machines to understand and generate text akin to human writing. This section provides a historical overview of their development and delves into the technical principles that underpin these state-of-the-art language models.

2.1. History and current status of large language models

Mentioning large language models inevitably brings us to the foundational field of Natural Language Processing (NLP). NLP is an interdisciplinary domain that intersects computer science, artificial intelligence, and linguistics. It focuses on developing computer systems that can effectively communicate using natural language. In the following, the author will discuss the historical development of the NLP field in three distinct periods.

In 1950, Alan Turing introduced the famous "Turing Test," which is considered the inception of NLP ideas. From the 1950s to the 1970s, during the era known as rule-based natural language processing, researchers attempted to mimic human cognitive language processes for natural language understanding. They heavily relied on manually crafted rules to process language [3]. However, the limitations of rule-based approaches, with their inability to cover all possible cases and the high cost of development and maintenance, posed challenges in handling complex languages.

From the 1970s to the early 21st century, with the rapid expansion of the internet and the abundance of language corpora, researchers witnessed significant progress in Natural Language Processing (NLP). Pioneered by researchers such as Jelinek at IBM Watson Laboratory, statistical-based approaches became prominent [4]. They established statistical language models based on the contextual characteristics of natural language, simplifying NLP problems into probabilistic ones. By employing these methods, the speech recognition rate was enhanced from 70% to 90%, marking a substantial breakthrough and transitioning NLP from the laboratory to practical applications. Techniques like N-gram models, Hidden Markov Models (HMM), Maximum Entropy models, and Support Vector Machines (SVM) demonstrated remarkable advancements in tasks such as machine translation, text classification, and information retrieval.

The third stage, beginning around 2010, saw the gradual integration of deep learning techniques into the field of NLP. Inspired by the success of deep learning in image and speech recognition, researchers started incorporating deep learning into NLP research. The introduction of the Word2Vec model in 2013 enabled the distributed representation of words, opening a new era for neural network-based NLP methods. In 2014, Google released the Seq2Seq model based on Long Short-Term Memory (LSTM), showcasing impressive performance in tasks like machine translation. The same year, Facebook developed a convolutional neural network-based model for text classification tasks.

DOI: 10.54254/2755-2721/35/20230399

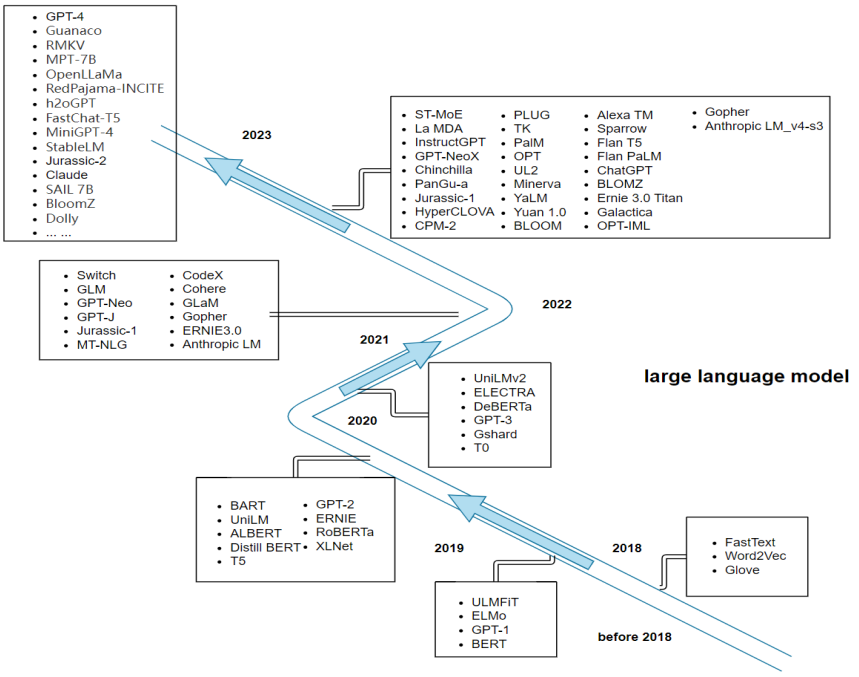

Figure 1. Refer to the picture in Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond and modify it (Photo/Picture credit: Original).

As depicted in the figure 1, the true breakthrough of large language models came with the publication of the paper "Attention is All You Need" in 2017 [5], which introduced the Transformer model for machine translation, proposed by the Google Brain team. Transformer, as the fourth major type of deep learning model after MLP, CNN, and RNN, stands out for its self-attention mechanism. It revolutionarily abandoned CNN and RNN in the sequence-to-sequence domain, relying solely on a simple network architecture with attention structures. This significantly accelerated the training process for sequence tasks and made it possible to create even larger models. The release of the BERT model marked the application of pre-training techniques in NLP [6]. BERT and its subsequent models like GPT-3, RoBERTa, etc., hav2.2 Mainstream architecture analysis of large language models - transformer modele achieved remarkable success in various NLP tasks.

2.2. Mainstream architecture analysis of large language models - transformer model

As stated earlier, the advent of the Transformer model revolutionized large language models, surpassing the previously dominant RNN-based approaches. So, what makes the Transformer model, with its crucial components of self-attention mechanism, multi-head attention, feed-forward neural networks, and positional encoding, superior [7].

2.2.1. Advantages of the transformer model. Firstly, the Transformer model showcases superior parallel efficiency compared to RNN models. The hidden state at a certain time step in RNNs depends on the output of the previous time step's hidden state, posing a barrier to parallel processing. In contrast, the Transformer model processes information from all contextual positions concurrently, minimizing the loss of information propagation.

Secondly, the Transformer model notably minimizes information propagation loss. RNN models, even enhanced variants like LSTM and GRU, may grapple with issues such as gradient explosion, vanishing gradients, or the forgetting of long-range dependencies when handling exceptionally long sequences. On the other hand, the Transformer model efficiently accesses information from all positions within a sequence, keeping the distance at one, mitigating these issues.

DOI: 10.54254/2755-2721/35/20230399

Lastly, the Transformer model proficiently amalgamates information from all positions. Unlike convolutional models, which usually consider smaller windows and require multiple convolution layers to integrate information from distant elements, the Transformer model adeptly collects information from all positions within a distance of one, enhancing the integration of information across the sequence.

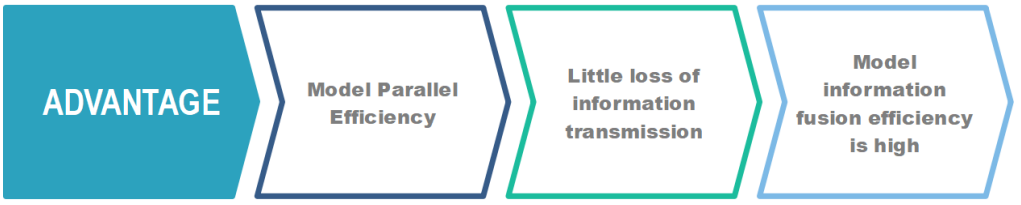

Figure 2. Advantages of the transformer model (Photo/Picture credit: Original).

As a result, the majority of models developed after 2017, such as the GPT series, BERT, RoBERTa, XLNet, and others, have adopted the Transformer architecture, making it the undisputed mainstream framework for large language models. As shown in Figure 2.

Moving forward, let's explore this attention-based neural network architecture in greater detail, focusing on two aspects: the self-attention mechanism and the diverse structures created by combining encoders and decoders.

2.2.2. Self-attention mechanism. The self-attention mechanism is an attention mechanism used for modeling sequential data. It computes attention weights by measuring the similarity between different positions in the input sequence. Given a position in the input sequence, the self-attention mechanism [8] calculates similarity scores between this position and other positions, normalizes the scores using the Softmax function, and obtains attention weights α (i, j) associated with other positions. These weights are then applied to the corresponding Value vectors (typically feature vectors from the input sequence) to produce the final attention output.

In large language models, the most common self-attention mechanisms are Scaled Dot-Product Attention and Multi-Head Attention.

In Scaled Dot-Product Attention, for each position i in the input sequence, similarity scores Score (i, j) are computed by taking the dot product of the Query vector and the Key vector of other positions j. The scores are then normalized using the Softmax function to obtain attention weights α (i, j). Finally, these weights are applied to the Value vectors to compute the attention output Ai for position i. The calculation formula is as follows:

$$Score(i,j) = Q_i \cdot k_j \tag{1}$$

$$\alpha(i,j) = \frac{\exp\left(\frac{Score(i,j)}{\sqrt{d}}\right)}{\Sigma j \exp\left(\frac{Score(i,j)}{\sqrt{d}}\right)} \tag{2}$$

$$A_i = \Sigma_j \, \alpha(i,j) \cdot V_j \tag{3}$$

In Multi-Head Attention, an extension of Scaled Dot-Product Attention, multiple independent attention heads are used, each with its own Query, Key, and Value vectors. This allows the model to learn different weights and compute attention from different perspectives on the input sequence. For each position i in the input sequence, let there be h attention heads, and the output of each head is represented as A_i^h. The final multi-head attention output A_i for position i is obtained by concatenating or linearly combining the outputs from all attention heads:

$$\mathrm{A_i} = \text{Concatenate}\left(\mathrm{A_i^1}, \mathrm{A_i^2}, \ldots, \mathrm{A_i^h}\right) \tag{4}$$

DOI: 10.54254/2755-2721/35/20230399

Presently, leading large language models, including GPT-3, BERT, RoBERTa, XLNet, Palm [9], and ChatGLM2 employ Multi-Head Attention as their self-attention mechanism. Multi-Head Attention allows the models to learn diverse representations and features, thereby enhancing their expressive capabilities and demonstrating superior performance in handling complex tasks and long sequences [10].

2.2.3. Different structures for combining encoders and decoders. As shown in Figure 3, similar to most seq2seq models, the Transformer architecture consists of an encoder and a decoder. Overall, LLM (Language Model) models can be categorized into three major types: Encoder-decoder Architecture, Causal Decoder Architecture, and Prefix Decoder Architecture [11].

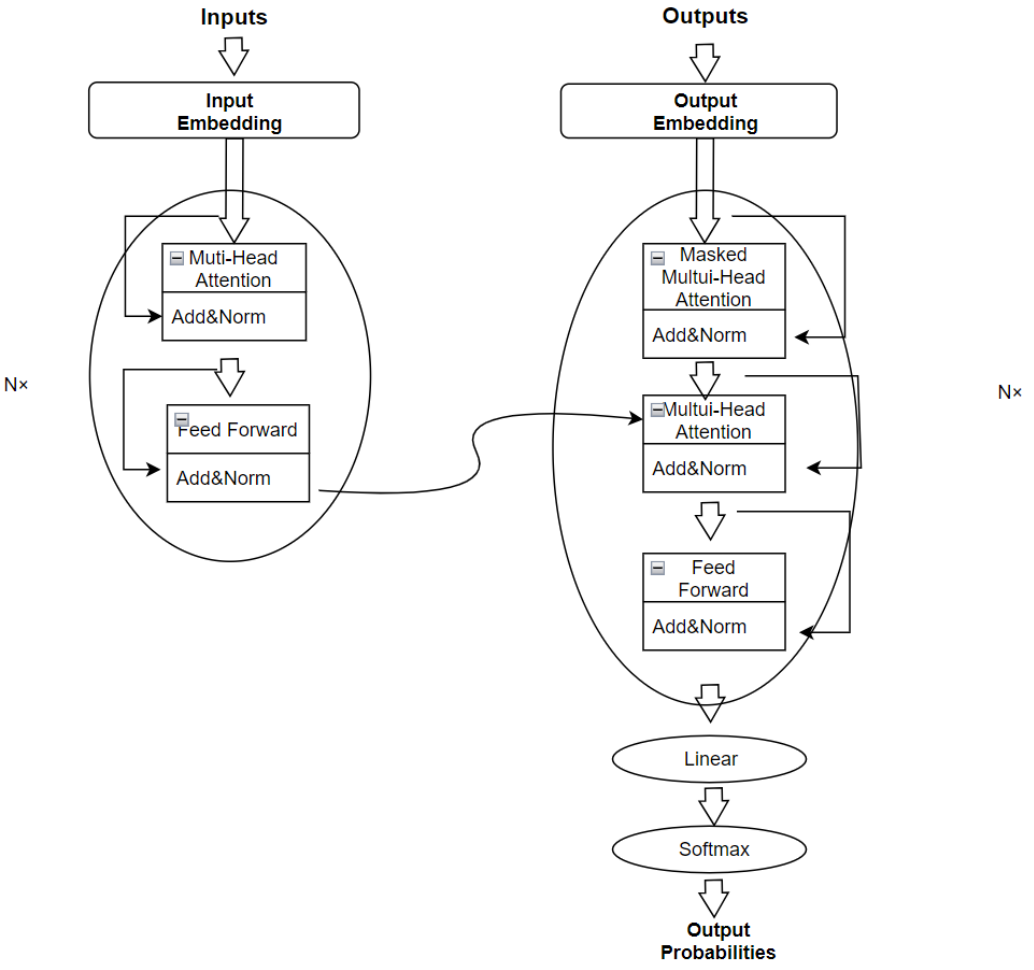

Figure 3. Structure diagram of Transformer model (caption centred) (Photo/Picture credit: Original).

The Encoder-decoder Architecture uses the most basic structure and was initially introduced by the Seq2Seq model to address sequence-to-sequence tasks, such as machine translation. It consists of an encoder and a decoder. The encoder is responsible for transforming the input sequence into a fixed-dimensional semantic representation, while the decoder uses this semantic representation to generate the output sequence. Within the encoder-decoder structure, self-attention mechanisms are commonly employed for sequence modeling, enabling the model to handle variable-length input and output sequences. This architecture has proven to be highly effective in various sequence-to-sequence tasks, such as text translation, text summarization, and dialogue generation. Prominent examples of large language models following this architecture include ELMo, BERT, RoBERTa, among others [12].

DOI: 10.54254/2755-2721/35/20230399

Currently, the most widely used architecture is the Causal Decoder, which is primarily employed for handling autoregressive generation tasks, where each element of the output sequence depends on previously generated elements. The Causal Decoder Architecture is an improvement over the Encoder-decoder structure, as it introduces an autoregressive mechanism in the decoder. This means that when generating the current element, the model only uses the elements generated before it. This ensures that the model does not have access to future information during the generation process, thereby preserving causality. The GPT series (e.g., GPT-3) is a typical example of models that use the Causal Decoder Architecture. These models generate text by sequentially producing words one by one, avoiding information leakage and enabling the generation of coherent and plausible text.

Compared to the Encoder-decoder models, Decoder-only models offer several advantages due to their simpler structure, faster training and inference speed, suitability for pure generation tasks, and advantages in decoder self-supervision. With the release of GPT-3 in 2021, Decoder-Only models have become a turning point and have gradually replaced Encoder-Decoder models, dominating the development of LLMs.

2.3. Large language model training methods - pre-training and fine-tuning

The concept of pretraining diverges from the traditional training approach based on backpropagation in neural networks, where network parameters are initialized randomly. Instead, pretraining involves training the model on specific tasks to obtain pretrained parameters. These pretrained parameters are then used to initialize the model before further fine-tuning. Pretraining is categorized under transfer learning [13].

The first-generation pretraining model emerged in 2013 with the introduction of word2vec, which provided word representations for training neural networks. However, it had limitations in effectively addressing the issue of polysemy in word embeddings. In 2018, ELMo marked the beginning of the second-generation pretraining language models, adopting the "pretraining + fine-tuning" paradigm. ELMo utilized bidirectional LSTMs as feature extractors, considering contextual word embeddings to better represent polysemous words [14]. Later, the more powerful Transformer architecture was applied to various subsequent pretraining language models like GPT and BERT, continually achieving state-of-the-art performance in natural language processing tasks [15].

In recent years, the success of PTMs (Pretrained Language Models) lies in integrating self-supervised learning with the Transformer architecture. Two highly influential PTMs, GPT and BERT, are based on the Transformer and utilize different models, namely autoregressive and autoencoder, respectively [16].

Autoregressive models predict preceding and subsequent words based on context. For instance, ELMo concatenates two autoregressive models, one running from left to right and the other from right to left, creating a bidirectional language model. Nevertheless, it still fundamentally belongs to the autoregressive model category.

Autoencoder models can be viewed as a denoising process, where pretraining involves predicting masked words based on context. The advantage of this model lies in its ability to utilize contextual information for predicted words. However, during the fine-tuning stage, masked words do not appear, leading to inconsistencies between the pretraining and fine-tuning stages due to the presence of [MASK] tokens.

During the pretraining phase, the model learns from an extensive and diverse dataset. Pretraining data can be broadly categorized into general data and specialized data. General data, such as web pages, books, and conversational texts, are commonly used due to their large scale, diversity, and accessibility. This stage enables the model to learn general language patterns and representations, enhancing its language modeling and generalization capabilities [17].

In the fine-tuning stage, the model is further trained on smaller and more specific datasets related to the target task or domain. This process combines the pretrained model's generalization ability with the specific requirements of the target task, resulting in improved performance. The fine-tuning process typically involves freezing certain parameters, updating top-level parameters, and adjusting others. Once fine-tuning is completed, the model can be deployed for specific tasks in practical applications. Notably,

DOI: 10.54254/2755-2721/35/20230399

Child-Tuning, proposed by Runxin Xu et al., updates a subset of parameters in large pretrained models by strategically masking gradients during the backward process, consistently achieving higher scores compared to regular fine-tuning (1.5-8.6 points) [18]. Additionally, Long Ouyang et al. fine-tuned GPT-3 using supervised learning and collected ranking data to further fine-tune the model through reinforcement learning with human feedback, resulting in the InstructGPT model. This model demonstrated improvements in realism and reduced generation of toxic outputs, while maintaining performance regression on public NLP datasets [19].

3. Application analysis of large language models

3.1. Overview of application areas

As depicted in the figure 4, large language models are currently widely utilized in diverse domains, including finance [20], [21], healthcare [22]-[24], education [25]-[27], law [28], geographical research [29], and literature [30]. The main application approaches encompass information extraction and integration, text generation, as well as analysis, prediction, and recommendation [31]. Moving forward, let's delve into their applications in the finance domain.

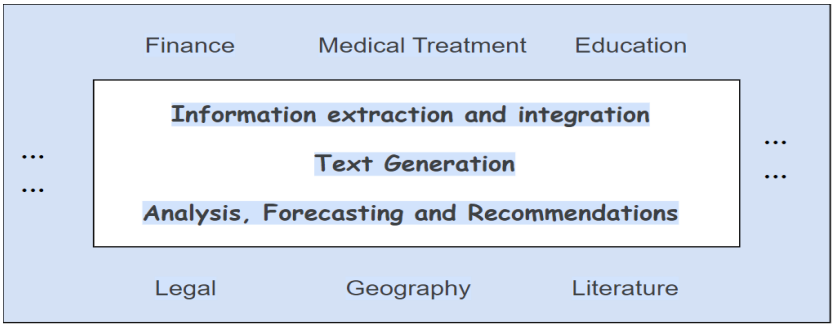

Figure 4. Main application methods and main application areas (Photo/Picture credit: Original).

3.2. Typical application cases

The financial industry, being a specialized field in managing financial commodities, stands out as one of the domains where large language models have achieved remarkable advancements.

Hongyang Yang pioneered the development of FinLLM, a FinTech language model capable of providing robot advice, algorithmic trading, and low-code development. They achieved this by leveraging automatic data management pipelines and lightweight low-rank adaptation techniques. Shijie Wu et al [32]. Constructed BloombergGPT, a financial large language model, by harnessing Bloomberg's extensive data sources, including 363 billion tokens from tokenized financial data and an additional 345 billion tokens from general datasets. FinBert, an open-source pretraining natural language processing (NLP) model, specifically trained on financial data, outperformed almost all other NLP technologies used for financial sentiment analysis [33]. Additionally, domain-specific financial large language models, such as the Chinese "Xuanyuan," are continuously under development.

Financial large language models serve as robot advisors, delivering personalized financial advice by analyzing investors' risk preferences, financial situations, and investment goals. These models recommend investment portfolios that align with their needs, thereby enhancing the accuracy and efficiency of investment decisions. They excel in analyzing financial market data and trends, providing trading signals for informed trading decisions. By integrating historical data, market trends, and risk assessments, these models assist institutions and investors in optimizing investment portfolios. They conduct sentiment analysis on social media, news, and public opinion data to better predict market trends. Furthermore, they aid financial institutions in achieving more effective risk management, including credit risk assessment, bankruptcy prediction, and corporate merger forecasts. Through the analysis of transaction data, these models identify potential fraudulent activities, thereby enhancing the security of financial transactions. ESG (Environmental, Social, and Governance) scoring has gained increasing

DOI: 10.54254/2755-2721/35/20230399

attention in the financial industry. Financial large language models can analyze data on companies and assets, providing investors with evaluations and rankings related to ESG performance, thereby promoting sustainable and socially responsible investments. They also contribute to enhancing financial literacy through financial education initiatives [33], continuously empowering the financial industry. These models offer more intelligent and efficient solutions to financial institutions and practitioners, driving innovation and development in the financial sector.

However, with the widespread application of large language models, a series of challenges arise, including ensuring high reliability and security in their deployment. These issues will be discussed in the next chapter.

4. Challenges and possible solutions

The prevailing challenges facing large language models can be categorized into: misinformation, ethical concerns, potential bias; data privacy issues; training cost challenges; and multi-modal application challenges. First, let's delve into misinformation, ethical concerns, and potential bias. ChatGPT often produces factually incorrect or biased outputs. Although this issue is intrinsic to generative AI models, its handling of the problem has been criticized [34]. As large language models find more applications, the repercussions of misinformation, ethical oversight, and biases, whether based on race, region, or other factors, intensify. The credibility of ChatGPT as a primary tool for healthcare education is compromised by its frequent inaccuracies, which learners might overlook [35]. Such issues not only hinder the broader adoption of these models but also deter users, including patients and practitioners, from trusting their responses. This, in turn, curtails the potential of automating healthcare processes. Researchers from institutions like OpenAI and the Stanford Institute for Human-Centered Artificial Intelligence have been proactive in addressing these issues [36]. Solutions discussed include leveraging metadata from online platforms, adopting encryption technologies for media authentication, utilizing diverse datasets, innovating anti-discrimination and bias-prevention algorithms, collaborating with a range of societal stakeholders, and advocating for relevant regulations.

Next, let's tackle the data privacy challenges. In 2023, two incidents of device information exposure and one of meeting content leakage occurred within a span of 20 days when Samsung permitted employees to utilize a certain large language model. Another notable incident involved a bug in the Redis open-source library, which triggered a malfunction and subsequent data leak of a large language model. Current technologies aimed at privacy protection are restricted by the fluid nature of privacy, the intricacies of defining private data, and the hurdles in pinpointing privacy informants. With added context, data becomes more dispersed, converting the concept of privacy protection from static to dynamic [37]. To address these concerns, H. Brown and team have suggested training language models exclusively on data that is designated for unrestricted public use, both now and in the future [38].

On the front of training costs, it's undeniable that they pose a significant barrier to the evolution of large language models. While advancements have been made through hardware improvements, model distillation, and accelerated training methodologies, the costs remain a substantial concern. As Professor Yoav Goldberg from Bar-Ilan University in Israel pointed out on GitHub, the "data-hungry" nature of these models makes it formidable to emulate the proficiency achieved in English for other languages, be it widely-spoken ones like German, French, Arabic, Chinese, or Hindi, or "low resource" languages prevalent in regions of Africa and the Philippines.

Lastly, the horizon looks promising for multi-modal applications. OpenAI's DALL·E, a variant of GPT-3 trained for text-to-image transformations, has made waves in the community. Similarly, Danny Britz and team have trained a robust multi-modal model, PaLM-E, capitalizing on comprehensive joint training spanning language, vision, and visual language domains at an unprecedented scale. Additionally, "Irene" integrates cutting-edge NLP and image recognition to aid in medical diagnostics [39]. These innovations pave the way for leaps in creativity, assisted production, and cross-domain integration.

DOI: 10.54254/2755-2721/35/20230399

5. Conclusion

Large language models have not only spurred academic advancements but have also seen extensive applications across diverse sectors such as finance, healthcare, biology, and education, imbuing these industries with unprecedented potential. However, as their integration grows more widespread, a myriad of ethical, economic, and other intrinsic challenges emerge, necessitating immediate and discerning solutions. Consequently, both scholars and industry professionals are intensely examining these models, looking for ways to alleviate these concerns and maximize their positive impact.

Research meticulously traces the historical development and core principles of large language models, with an emphasis on their ramifications in the financial sector. The challenges that have arisen during their evolution are addressed, and well-founded solutions are proposed. Through this in-depth exploration, a vivid tableau of insights emerges, shedding light on the past, present, and potential futures of large language models. Looking forward, it's clear that as technological innovation quickens and research becomes more profound, large language models will continue to play a central role in NLP's unfolding story. There's optimism for upcoming enhancements to tackle current challenges, catalyzing even more groundbreaking innovations for society. Yet, such optimism should be paired with caution, constantly assessing potential setbacks and risks to ensure that technology genuinely serves humanity's greater good. With persistent research and thoughtful exploration, there's confidence that large language models can pave a more promising and advantageous path for the global future.

References

[1] Norris, J. R. (1998). Markov chains (No. 2). Cambridge university press.
[2] Markov, A. (2006). An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. Science in Context, 19(4), 591-600.
[3] Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine, 9(2), 48-57.
[4] Jarrinik, A. (1999). "From Watergate to Monica Lewinsky": Presentation at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Conference Proceedings, 1999, 4-7.
[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[7] Bommasani, R. , Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
[8] Lei, S., Yi, W., Ying, C. , & Ruibin, W. (2020). Review of attention mechanism in natural language processing. Data analysis and knowledge discovery, 4(5), 1-14.
[9] Chowdhery, A. , Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.
[10] Shazeer, N. (2019). Fast transformer decoding: One write-head is all you need. arXiv preprint arXiv: 1911.02150.
[11] Zhao, W. X., Zhou, K., Li, J., Tang, T. , Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
[12] Yang, J., Jin, H. , Tang, R., Han, X., Feng, Q., Jiang, H., ... & Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. arXiv preprint arXiv:2304.13712.
[13] Zhang, J., Huang, J. , Jin, S., & Lu, S. (2023). Vision-language models for vision tasks: A survey. arXiv preprint arXiv:2304.00685.
[14] Peters, M. E., Neumann, M. , Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365. Retrieved from https://arxiv.org/abs/1802.05365.
[15] Radford, A., Narasimhan, K., Salimans, T. , & Sutskever, I. (2018). Improving language understanding by generative pre-training.

DOI: 10.54254/2755-2721/35/20230399

[16] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X. , Huo, Y., ... & Zhu, J. (2021). Pre-trained models: Past, present and future. AI Open, 2, 225-250.
[17] Zhao, W. X., Zhou, K., Li, J. , Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
[18] Xu, R., Luo, F., Zhang, Z., Tan, C., Chang, B., Huang, S., & Huang, F. (2021). Raise a child in large language model: Towards effective and generalizable fine-tuning. arXiv preprint arXiv:2109.05687.
[19] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.
[20] Yang, H., Liu, X. Y. , & Wang, C. D. (2023). FinGPT: Open-Source Financial Large Language Models. arXiv preprint arXiv:2306.06031.
[21] Dowling, M. , & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. Finance Research Letters, 53, 103662.
[22] Xue, K., Zhou, Y., Ma, Z., Ruan, T., Zhang, H., & He, P. (2019, November). Fine-tuning BERT for joint entity and relation extraction in Chinese medical text. In 2019 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 892-897). IEEE.
[23] Wang, S., Zhao, Z., Ouyang, X. , Wang, Q., & Shen, D. (2023). Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257.
[24] López-Úbeda, Pilar, Teodoro Martín-Noguerol, and Antonio Luna. "Radiology in the era of large language models: the near and the dark side of the moon." European Radiology (2023): 1-3.
[25] Kasneci, E. , Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.
[26] Gilson, A. , Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Medical Education, 9(1), e45312.
[27] Kung, T. H. , Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS digital health, 2(2), e0000198.
[28] Xiao, C., Hu, X., Liu, Z., Tu, C. , & Sun, M. (2021). Lawformer: A pre-trained language model for chinese legal long documents. AI Open, 2, 79-84.
[29] Roberts, J. , Lüddecke, T., Das, S., Han, K., & Albanie, S. (2023). GPT4GEO: How a Language Model Sees the World's Geography. arXiv preprint arXiv:2306.00020.
[30] de la Rosa, J., Pozo, Á. P., Ros, S., & González-Blanco, E. (2023). ALBERTI, a Multilingual Domain Specific Language Model for Poetry Analysis. arXiv preprint arXiv:2307.01387.
[31] Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., ... & Zhao, L. (2023). Beyond One-Model-Fits-All: A Survey of Domain Specialization for Large Language Models. arXiv preprint arXiv:2305.18703.
[32] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.
[33] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.
[34] Zhou, J., Ke, P., Qiu, X., Huang, M., & Zhang, J. (2023). ChatGPT: Potential, prospects, and limitations. Frontiers of Information Technology & Electronic Engineering, 1-6.
[35] Thirunavukarasu, A. J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., El Mukashfi, M., & Shah, S. (2023). Trialling a large language model (ChatGPT) in general practice with the Applied Knowledge Test: observational study demonstrating opportunities and limitations in primary care. JMIR Medical Education, 9(1), e46599.

DOI: 10.54254/2755-2721/35/20230399

[36] Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503.
[37] Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022, June). What does it mean for a language model to preserve privacy?. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2280-2292).
[38] Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022, June). What does it mean for a language model to preserve privacy?. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 2280-2292).
[39] Zhou, H. Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., ... & Li, W. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering, 1-13.