WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Recent Advances in Large Language Models: An Upshot

[1]Rithika

[1]Student, Master of Computer Applications, School of CS & IT, Jain (Deemed-To-Be-University), Bangalore, India,
[1]rithusuvarna07@gmail.com

DOI: https://doi.org/10.55248/gengpi.5.0624.1403

A B S T R A C T

This article discusses the recent applications of large language models (LLMs) in different sectors like natural language processing (NLP), healthcare, finance, and education. We research how LLMs affect different tasks like text generation, machine translation, sentiment analysis, and question-answering systems in the field of natural language processing. An analysis is conducted on the advancements in NLP driven by models like GPT-4, BERT, and T5, demonstrating their capacity to understand and generate text that mimics human language. In the healthcare industry, research is conducted on the utilization of LLMs to improve diagnostic precision, predict patient outcomes, and advance personalized healthcare. Specific instances of singular cases, such as using LLMs to analyze electronic health records and assisting medical research through automated literature reviews, are given. Within the financial sector, there is a strong emphasis on incorporating LLMs and analyzing their influence on algorithmic trading, risk assessment, fraud identification, and automation of customer service. We provide examples of how LLMs are used to parse financial news, generate investment ideas, and manage large datasets for improved decision-making. In the education sector, we explore ways to integrate LLMs into personalized language, automated grading, and tutoring platforms. The article discusses how LLMs can create personalized language environments to address students' individual needs by offering tailored support and resources. We investigate the benefits of LLMs, such as their efficiency in managing and generating large volumes of data, improving accuracy in predictions and evaluations, and simplifying complex tasks with automation. Moreover, we address the obstacles presented by these challenges, including worries about data privacy, ethical conflicts, and the computational resources required for training and deploying such models. The review of literature is comprehensive, detailing significant research papers, their authors, and key findings. This assessment includes significant research from scholars such as Vaswani et al. (2017) regarding the Transformer architecture.

Keywords: Large Language Models, Natural Language Processing, Healthcare, Finance, Education, Machine Language

INTRODUCTION:

In recent years, there have been notable advancements in large language models, especially those utilizing deep language and neural network designs. These architectures, like GPT-3, BERT, and others, have demonstrated impressive abilities in comprehending and producing text that resembles human language [1] [2]. Created with large sets of data and advanced algorithms, these models have reached new heights in performance for different natural language processing (NLP) tasks. As an example, GPT-3 has established new standards in text generation, translation, and conversational AI by utilizing its 175 billion parameters to exhibit a thorough comprehension of context and semantics [3].

In the field of natural language processing, these extensive language models have drastically changed how machines engage with human language. BERT, created by Google, allows for better text comprehension in search queries by considering the meaning of words in their context [16].

This bidirectional approach allows BERT to understand nuances in language, ensuing in enormous improvements in obligations which includes sentiment analysis, named entity reputation, and question answering. similarly, OpenAI's GPT-three has been leveraged to expand superior chatbots and virtual assistants which could have significant and coherent conversations with customers, improving consumer experience to feel greater instinctive and genuine.

The healthcare sector has greatly benefited from the use of large language models. These models have been utilized to improve diagnostic tools, forecast patient outcomes, and simplify administrative procedures. [8] LLMs have the ability to review extensive medical literature and electronic health records to help identify rare diseases or recommend treatments using the most current research. Furthermore, LLMs have the capability to aid in the advancement of personalized medicine through the analysis of patient information for predicting individual reactions to treatments, ultimately enhancing the accuracy and efficacy of medical interventions.

In the field of finance, advanced language models are changing conventional methods by allowing for more complex analysis and automation [9]. Financial institutions utilize LLMs for various purposes.

APPLICATIONS OF LLM IN VARIOUS DOMAINS:

Large language models have transformed natural language processing (NLP) through greatly improving obligations like language translation, text summarization, sentiment analysis, and question-answering structures. models which includes BERT and GPT-three have received sizable popularity and reward for his or her impressive improvements in these fields. BERT has transformed tasks such as sentiment analysis and named entity recognition by understanding context from both directions, while GPT-3's large scale allows it to create text resembling human writing and provide contextual responses to questions. These models have greatly changed the way machines comprehend and create language, expanding the scope of NLP studies and uses.

In the field of healthcare, big language models have become important resources for forecasting disease outbreaks, identifying medical issues, and tailoring treatment plans. Specialized models such as BioBERT and ClinicalBERT are created to comprehend biomedical literature and clinical notes, resulting in notable progress in biomedical NLP tasks. These models use large quantities of medical data to extract valuable insights, helping healthcare professionals make decisions and enhance patient results. By examining electronic health records and medical research papers, these models play a role in advancing medical science and improving patient care.

The finance sector has also adopted extensive language models for examining financial statements, forecasting market patterns, and identifying fraudulent behaviors. Specialized models such as FinBERT are designed to understand financial communications in order to offer more precise insights and predictions. Through reading huge portions of economic information, those fashions assist monetary analysts in making properly-knowledgeable selections, coping with risks, and spotting investment potentialities. additionally, they have got a critical feature in figuring out fraudulent sports by way of examining transaction styles and detecting suspicious behavior. within the subject of training, huge language models play a key function in developing smart tutoring systems, automating grading, and delivering personalized language possibilities.

LITERATURE REVIEW:

The Transformer model, presented by Vaswani et al. “Attention is All you Need” [1] in 2017, has become essential for various extensive language models such as BERT and GPT. The Transformer structure transformed the field of natural language processing by getting rid of recurrent neural networks (RNNs) and using self-attention mechanisms to handle input data. This new development enables the model to take in the complete sequence context all at once, instead of one by one, significantly enhancing the efficiency and effectiveness of both training and inference. The Transformer is highly suitable for tasks like language translation, text generation, and understanding due to its ability to capture intricate dependencies in the data through the self-attention mechanism, particularly useful for long-range contexts. This advancement paved the way for future models such as BERT, which utilizes bidirectional context for detailed language comprehension, and GPT, which thrives in generating text autonomously, expanding the possibilities of deep language in NLP.

BERT (Bidirectional Encoder Representations from Transformers), a version that driven the limits of herbal language processing (NLP) by using excelling in eleven responsibilities due to its particular bidirectional schooling technique changed into introduced by Bert: Pre-training of DEEP Bidirectional Transformers for Language know-how [2] contrary to beyond fashions, BERT's bidirectional schooling allows it to analyze context from both guidelines on the equal time, rather than simply in a single course. BERT's notable performance in tasks like question answering, named entity reputation, and sentiment analysis is a end result of its thorough draw close of context. via utilizing large pre-schooling on massive textual content datasets and then satisfactory-tuning on precise responsibilities, BERT has mounted new requirements in NLP, showcasing its adaptability and efficiency in grasping the subtleties of human language.

“Language fashions are Few-Shot learners” [3] defined OpenAI's advent, GPT-three, that's an outstanding language version with 175 billion parameters, rating as certainly one of the most important and most mighty models in its class. The extensive variety of parameters in GPT-3 lets in it to understand and bring text that intently resembles human language with extraordinary precision and corporation. One of its key characteristics is its capacity to carry out a variety of tasks with little training, sometimes only needing a few examples or a prompt to generate relevant responses within the context. GPT-3 is incredibly skilled in textual content generation, translation, summarization, and query-answering, showcasing an awesome ability to modify to diverse contexts and programs without the need for specialised quality-tuning. the flexibility and user-friendliness of GPT-three make it a valuable device in a huge variety of fields, from customer service and content material advent to programming aid and research, displaying its capacity to convert how we interact with and make use of AI generation.

XLNet: Generalized Autoregressive Pretraining for Language Understanding [4] proposed the XLNet version, which included the strengths of each autoregression and autoencoding fashions, accomplished higher overall performance than BERT. XLNet captures bidirectional context without masking by using a permutation-based training approach, unlike BERT which relies only on masked language modeling for understanding word context. This approach enables XLNet to effectively grasp relationships between words, leading to better outcomes in different natural language processing assignments. XLNet blends the predictive power of autoregressive models with the contextual understanding of bidirectional autoencoding models, resulting in a deeper grasp of language and top-notch performance in tasks like question answering, sentiment analysis, and text classification.

Improved BERT's effectiveness has been attained through training with bigger batches and increased data, enabling the model to comprehend and extrapolate from a wider array of language contexts in RoBERTa: A Robustly Optimized BERT Pretraining Approach [5]. Using bigger batch sizes in training results in better optimization by making gradient estimates more steady, which leads to faster convergence and higher model accuracy. Furthermore, by including broader and more varied datasets, the model can better grasp a wider range of language patterns and subtleties, resulting in

enhanced performance across different natural language processing tasks like text categorization, identifying named entities, and answering questions. This method doesn't just increase the model's resilience but also enhances its capacity to deal with intricate language structures and infer meaning with greater accuracy.

ALBERT: A Lite BERT for Self-supervised Language of Language Representations [6] proposed current improvements in large language models have concentrated on decreasing model size and enhancing training speed while maintaining performance. Methods like knowledge distillation, model pruning, and quantization have been crucial in reaching these objectives. Knowledge distillation consists of teaching smaller models to imitate the actions of larger models, successfully passing on knowledge while upholding high accuracy. Model pruning decreases the size of neural networks by eliminating unnecessary parameters, consequently simplifying the model with minimal impact on performance. Quantization decreases the accuracy of the model's weights, leading to reduced memory usage and faster computation. Combining these techniques allows for the implementation of effective, robust models that can be trained and executed more quickly, increasing their utility and feasibility for real-life uses, particularly on devices with restricted computational capabilities.

BioBERT [7] the adaptation of BERT for biomedical texts, has greatly enhanced performance in various biomedical natural language processing (NLP) tasks. BioBERT improves the model's comprehension of the distinctive terminology and context in the biomedical field by training it on a large collection of biomedical literature, such as PubMed abstracts and PMC full-text articles. BioBERT has been trained in a specific way that allows it to excel in various tasks, such as recognizing named entities in the biomedical field, like genes, diseases, and drugs, with improved precision. Furthermore, BioBERT stands out in relation extraction, analyzing connections between biomedical entities, and in question answering, giving accurate and contextually appropriate responses to queries using biomedical texts. These enhancements make it easier to find and analyze information in biomedical research and healthcare applications, showing the advantages of customizing large language models for specific domains.

ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission [8] utilized BERT on medical records, improving forecasts of rehospitalizations by utilizing its capacity to comprehend the context and intricacies of medical terminology. BERT can detect subtle indicators and patterns that traditional models may miss by analyzing unstructured text data from electronic health records. This permits for a higher forecast of which patients can also have a greater chance of being readmitted, permitting healthcare companies to behave quicker with particular care plans and assets to enhance patient outcomes and decrease healthcare charges.

FinBERT: A Pretrained Language Model for Financial Communication [9] developed a specialised model of BERT for studying economic text called FinBERT, this version is designed to efficiently interpret the intricacies and particular terminology of the economic industry.. Through extensive training on financial datasets, FinBERT gains proficiency in comprehending and interpreting financial terminology, jargon, and context-specific nuances present in materials such as earnings calls, market reports, and regulatory filings. This optimization helps FinBERT stand out in jobs like analyzing market news sentiment, identifying important financial entities, and forecasting market trends using text data. As a result, FinBERT offers financial experts a robust resource for obtaining practical information, improving decision-making procedures, and streamlining the assessment of extensive amounts of financial text with excellent precision and importance.

Customized specifically for scientific literature, this specialized version called SciBERT aims to enhance accuracy in tasks related to scientific texts in SciBERT: A Pretrained Language Model for Scientific Text [10]. Through pre-training on a extensive range of scientific publications from different fields, SciBERT develops a profound grasp of the distinct vocabulary, terminologies, and intricate structures commonly found in academic writing. This specialized pre-training for specific domains improves the performance of SciBERT in tasks like extracting information, recognizing entities, and extracting relations in scientific papers. As a result, SciBERT helps with more precise and efficient literature reviews, automatic metadata creation, and uncovering new research findings, thereby assisting researchers in handling and exploring the expanding scientific knowledge base more effectively.

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [11] created a compressed version of BERT that retains 97% of its language understanding capabilities, this optimized model, known as TinyBERT, employs advanced model compression techniques such as knowledge distillation, weight pruning, and quantization to significantly reduce its size and computational requirements. Despite the reduction in parameters and memory footprint, TinyBERT maintains nearly the same level of performance as the original BERT in various natural language processing tasks. This balance of efficiency and effectiveness makes TinyBERT especially treasured for deployment in aid-confined environments, which include cell devices and side computing applications, where computational power and reminiscence are restrained however high language knowledge accuracy is still required.

GPT-2: Better Language Models and Their Implications [12] showcased GPT-2's capacity to create coherent text paragraphs, showcasing the model's advanced skills in herbal language generation via producing applicable and human-like textual content. The remarkable outcomes led to extensive debates surrounding AI ethics, specifically regarding the risk of abuse in generating deceptive content, deepfakes, and automated disinformation efforts. Concerns were brought up regarding authorship and originality, as well as the necessity for regulations and guidelines to guarantee responsible utilization of these potent technologies. The discussion also covered the wider societal effects, such as the possibility of AI-generated content shaping public beliefs and the importance of creating strong systems to identify and address unethical uses of AI.

T5: Exploring the Limits of Transfer Language with a Unified Text-to-Text Transformer [13] demonstrated a groundbreaking approach by redefining how different natural language processing tasks are tackled and carried out, demonstrating the adaptability of the T5 model through the framing of all NLP tasks as text-to-text transformations. Through converting various tasks' inputs and outputs into a unified text-to-text format, T5 simplifies the modeling process and enables one model to efficiently handle a broad range of language tasks, including translation, summarization, question answering, and classification. This standardized structure makes training easier and also utilizes transfer language, which improves overall model performance by

using knowledge gained from one task to help with others. The T5 model's strength in being able to apply to various NLP tasks highlights its resilience and adaptability, establishing it as a potent asset in enhancing language model capabilities.

Megatron-LM [14] introduced techniques for effectively training very large models through model parallelism, where a huge neural network is divided into smaller components that can be processed simultaneously across multiple GPUs or computational nodes. This method solves the memory and computational limitations often faced during the training of massive models by distributing the tasks, allowing the training of models with billions of parameters that would be impossible on a single device. Methods like pipeline parallelism, which processes various layers of the network in phases, and tensor slicing, which splits and distributes tensors, improve resource efficiency and decrease training duration. The progress in model parallelism speeds up training and improves scalability, allowing for the creation of more advanced deep language models that push the limits in areas such as natural language processing and computer vision.

Electra [15] proposed a pretraining method that is more efficient in terms of samples and surpasses BERT in performance, using techniques like curriculum language, dynamic masking, and optimized tokenization to improve the language process, all while reducing computational costs. This method decreases the training iterations needed by selecting the most informative samples and gradually enhancing the complexity of the training data, thus maintaining or enhancing performance. Moreover, it utilizes a more strategic allocation of computer resources, leading to quicker convergence and lower energy usage. This effectiveness speeds up the creation of strong language models and also improves the training process accessibility and environmental sustainability, allowing wider use and advancement in natural language processing.

DeBERTa [16] improved BERT's effectiveness with the implementation of disentangled attention mechanisms, which divide attention heads to concentrate on various elements of the input information, enabling the model to more effectively grasp and depict intricate connections within the text. By unraveling the focus, the model can better regulate how attention is spread across different linguistic aspects like syntax, semantics, and context. This leads to a more detailed comprehension and creation of written content, resulting in enhancements in activities such as machine translation, text summarization, and question answering. The untangled attention mechanisms improve the model's explainability and decrease the computational burden, making the upgraded BERT more efficient and scalable for various NLP tasks.

GPT-Neo [17] developed an open-source model akin to GPT-3, this project utilizes the Mesh-TensorFlow framework to improve scalability and efficiency when training large neural networks. Mesh-TensorFlow helps spread out the training process over several GPUs or TPUs, allowing for large model sizes and huge datasets, similar to those utilized in GPT-3. This open-source model seeks to make advanced language modeling capabilities more accessible by offering a powerful tool for natural language processing tasks like text generation, translation, and summarization to the research and developer community. Through the use of Mesh-TensorFlow, the model is able to achieve top-notch results, all the while ensuring it remains adaptable and user-friendly, encouraging creativity and teamwork within the realm of artificial intelligence.

RealToxicityPrompts [18] analyzed and recommended strategies to decrease the generation of harmful language in large language models, this research focuses on alleviating the serious problem of toxic outputs that may occur in models such as GPT-3. The study entails a thorough investigation into the origins and varieties of harmful language, including hate speech, prejudiced remarks, and unsuitable material, that may be generated unintentionally by these models. Suggested methods to address the issue include improving the training data to remove harmful content, using real-time content filtering during content generation, and incorporating post-processing methods to identify and fix offensive results. Moreover, the research delves into utilizing RLHF to teach models to better adhere to ethical standards and meet user expectations. The aim is to improve the safety and dependability of big language models by utilizing these techniques, so they can have a positive impact on user interactions and reduce the dissemination of harmful language.

DAPT and TAPT [19] presented the methods to improve performance on specific tasks by customizing the pretraining of large language models with data from fields like medicine, law, or finance. Through the use of a carefully selected collection of domain-specific texts, these techniques guarantee that the language models gain a more thorough comprehension of the specialized vocabulary, setting, and distinctive language structures specific to the chosen area. This particular initial training greatly improves the models' precision and efficiency when carrying out responsibilities like identifying entities, categorizing text, and analyzing sentiment within the given field. As a result, domain-specific pretrained models produce more accurate and dependable results, enhancing the overall effectiveness and relevance of language models for tackling specialized challenges in various industries.

CTRL [20] developed a model that enables a higher level of control and guidance in text generation, this improved version boosts the features of traditional language models by integrating mechanisms that empower users to impact the direction, style, and content of the text produced. This model permits accurate modifications to tone, specificity, and thematic focus by incorporating more input parameters or conditioning variables, ensuring that the output meets user intentions and requirements closely. This regulated creation is especially beneficial in areas like automated content generation, where it is essential to uphold consistency with brand tone or adhere to specific story formats. Additionally, it aids in generating more precise and situationally relevant answers in conversational agents, ultimately enhancing user engagements and contentment.

The summarization of all the papers discussed are given in Table-1:

Table-1

Index	**Paper Title**	**Authors**	**Year**	**Summary**
1	“Attention is All You Need”	Vaswani et al.	2017	The Transformer model served as the basis for numerous major language models such as BERT and GPT.

2	“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”	Devlin et al.	2018	Introduced was BERT, a model that enhanced the current best performance in 11 NLP tasks by utilizing bidirectional training.
3	“Language Models are Few-Shot Learners”	Brown et al.	2020	GPT-3, a model containing 175 billion parameters, is known for its ability to tackle diverse tasks with little need for training.
4	“XLNet: Generalized Autoregressive Pretraining for Language Understanding”	Yang et al.	2019	Suggested XLNet surpassed BERT by combining the strengths of autoregressive and autoencoding models.
5	“RoBERTa: A Robustly Optimized BERT Pretraining Approach”	Liu et al.	2019	Improved BERT's effectiveness through training with bigger batches and increased data.
6	“ALBERT: A Lite BERT For Self-supervised Language of Language Representations”	Lan et al.	2019	Decreased model size and accelerated training while maintaining performance.
7	“BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining”	Lee et al.	2020	Modified BERT for medical texts, enhancing results in medical NLP assignments
8	“ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission”	Huang et al.	2019	Utilized BERT for analyzing clinical notes, improving accuracy in predicting hospital readmissions
9	“FinBERT: A Pretrained Language Model for Financial Communications”	Yang et al.	2020	Created a specialized version of BERT specifically designed for analyzing financial texts
10	“SciBERT: A Pretrained Language Model for Scientific Text”	Beltagy et al.	2019	Customizing BERT for academic research papers to enhance accuracy in scientific text-related tasks
11	“DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”	Sanh et al.	2019	Developed a condensed form of BERT that maintains 97% of its linguistic comprehension abilities..
12	“GPT-2: Better Language Models and Their Implications”	Radford et al.	2019	Demonstrated the capability of GPT-2 in producing cohesive text passages, initiating conversations about the ethics of AI.
13	“T5: Exploring the Limits of Transfer Language with a Unified Text-to-Text Transformer”	Raffel et al.	2020	Showcased the flexibility of the T5 model by presenting every NLP assignment as conversions between text inputs and outputs.
14	“Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism”	Shoeybi et al.	2019	Presented methods for effectively training extremely large models through the use of model parallelism.
15	“Electra: Pre-training Text Encoders as Discriminators Rather Than Generators”	Clark et al.	2020	Suggested a pretraining technique that is more effective in terms of sample efficiency and requires less computational resources than BERT

16	"DeBERTa: Decoding-enhanced BERT with Disentangled Attention"	He et al.	2020	Improved BERT's efficiency through the incorporation of disentangled attention mechanisms.
17	"GPT-Neo: Large Scale Autoregressive Language Modelling with Mesh-Tensorflow"	Black et al.	2021	Created an open-source model resembling GPT-3, utilizing the Mesh-Tensorflow framework.
18	"RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models"	Gehman et al.	2020	Examined and suggested ways to reduce the production of harmful language in extensive language models.
19	"DAPT and TAPT: Domain-Adaptive Pretraining and Task-Adaptive Pretraining for Biomedical Domain"	Gururangan et al.	2020	Introduced techniques for pretraining in specific domains to improve results on specialized assignments.
20	"CTRL: A Conditional Transformer Language Model for Controllable Generation"	Keskar et al.	2019	Created a framework enabling better regulated and directed text production

CHALLENGES AND FUTURE DIRECTIONS:

Large language models face various difficulties that need to be addressed even though they have impressive achievements. Important questions about the societal impact of these models are raised by ethical issues including data privacy, algorithmic bias, and potential misuse. Dealing with these issues involves taking into account ethical guidelines and regulations to guarantee the responsible and ethical deployment of these models. Additionally, the significant amount of computational power needed to train and use big models creates practical difficulties, restricting smaller research teams and organizations with fewer resources from having access. Future research is focused on creating better training methods and model structures to lower costs without sacrificing results, making advanced AI technology more accessible.

Another important issue is the possibility of prejudiced results from extensive machine language models, which may continue and magnify current societal prejudices found in the data used for training. Researchers are trying to reduce the risk by investigating ways to identify and lessen biases in models and datasets, while also working on strategies to enhance the interpretability and transparency of models. Improving the transparency of these models helps stakeholders comprehend decision-making processes, recognize biases, and take action as needed. Furthermore, initiatives are in place to broaden datasets and guarantee sufficient inclusion of various demographic categories, reducing the chance of reinforcing prejudices in AI systems. In general, upcoming studies in the area of extensive language models seek to tackle these obstacles and push forward the creation of more effective, understandable, and unbiased AI systems for a variety of uses.

CONCLUSION:

Machine language systems' capabilities have been significantly enhanced in various fields thanks to the use of large language models. Their flexibility and capabilities are showcased in natural language processing, healthcare, finance, and education, where they have transformed activities like language comprehension, medical diagnosis, financial analysis, and customized language. Continual research and development initiatives seek to enhance these models by addressing existing challenges such as ethical issues, computing demands, and biases, as well as discovering novel uses and revealing unforeseen opportunities. As big language models advance, they have the potential to inspire change and innovation in various fields, shaping the future of AI and machine language.

References:

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. arXiv preprint arXiv:1706.03762.
2. 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.
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.
4. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237.

5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
6. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Language of Language Representations. arXiv preprint arXiv:1909.11942
7. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinformatics, 36(4), 1234-1240.
8. Huang, K., Altosaar, J., Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv preprint arXiv:1904.05342.
9. Yang, Y., Zhang, Y., Chen, X., & Zhang, X. (2020). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:2006.08097.
10. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. arXiv preprint arXiv:1903.10676.
11. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
12. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). GPT-2: Better Language Models and Their Implications. arXiv preprint arXiv:1906.08781.
13. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). T5: Exploring the Limits of Transfer Language with a Unified Text-to-Text Transformer. arXiv preprint arXiv:1910.10683.
14. Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv preprint arXiv:1909.08053.
15. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv preprint arXiv:2003.10555.
16. He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv preprint arXiv:2006.03654.
17. Black, S., Leo, G., Wang, P., Leahy, C., Biderman, S., & Gao, L. (2021). GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. arXiv preprint arXiv:2101.00027.
18. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv preprint arXiv:2009.11462.
19. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). DAPT and TAPT: Domain-Adaptive Pretraining and Task-Adaptive Pretraining for Biomedical Domain. arXiv preprint arXiv:2004.10964.
20. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A Conditional Transformer Language Model for Controllable Generation. arXiv preprint arXiv:1909.05858.