Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/133/2025.20608

Enhancing the Performance and Applicability of AI Large Language Models: Strategies and Improvements

Pengfei Shen[1,a,*]

[1]*Institute of School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, The United States*
a. spengfeiunc@gmail.com
**corresponding author*

Abstract: This work examines the methods of optimizing Large Language Models, focusing on their applications in present-day implementations of Artificial Intelligence across sectors such as Natural Language Processing, customer support, and text creation. As a result, prior models have restrictions when it comes to processing the more complex linguistic features, proper usage of CPU and those models are not adhering the ethical practices. This research responds to these challenges by exploring advancements in model designomal structure and update methods, together with categorical responsibility frameworks. Measuring the efficiency of large language models and the potential for their optimization, the work also uses the analysis of user feedback. The research shows that small changes in model design and training strategies have high impact on model performance and interpretability in various language tasks. These findings help to advance the creation of improved and more reliable language models with optimised ethical frameworks and utility of resources. Areas for future work are presented, such as integration of more complex multi-modal data and decreasing the amount of bias in model outputs.

Keywords: AI Large models, Natural Language Processing, Machine Learning, Model Architecture, Ethical AI

1. Introduction

The capacity of large language models to perform tasks in the healthcare systems is immense. Today, the deployment of large language models to perform human-like tasks spans every system of humanity. We are bombarded by new version releases of large language models by firms like OpenAI, Google, Microsoft, and X daily in sectors like education, and research. However, in sectors such as healthcare, where people's lives are at risk, there should be stricter regulations on the release of large language models. A key driver for these regulations is the ethical framework, specific to healthcare.

Fine-tuning the model's performance on context-related tasks can also help to enhance the relevance and accuracy of large language models in key domains such as health care. For instance, the dataset used has specific terminology and vocabulary for healthcare staff which, if non-generic LLMs are exposed to frequently, they can be adapted to deliver specific tasks accurately: predict readmission rates based on clinical diagnosis and outcomes. Also, by understanding sector-specific vocabulary, it is possible to train large language models on a phrase which will later enhance how the LLM parses the context vocabulary.

DOI: 10.54254/2755-2721/133/2025.20608

Using LLMs it is possible to teach a system to read and write like a human, thus enhancing artificial intelligence. These models include OpenAI's GPT-4, and Google BERT among others, and its functionality is anchored on embeddings derived from transformer models. Due to possessing large datasets and training, LLMs are capable to do tasks that were deemed human privilege, such as translation, summarization, and writing. This leap has been used in scenarios such as Automated Customer Interaction where LLMs are applied to answer tricky questions from a user and in Virtual Personal Assistants where they enable more intelligent and more efficient natural interaction between the human and the machine. The improvement of LLMs has been occurring at a remarkable pace and the potential scenarios across industries have expanded enormously to introduce breakthrough levels of automation, productivity, and experience.

Although the previous years' literature has strongly concentrated on the way of enhancing model size and the number of parameters, which has beneficial effects in terms of performance, this strategy gives a relatively low level of effect in contemporary years. Amplifying the number of LLMs does not result in improved capabilities in handling complex languages and it is not effective thus. Bigger models are computationally more demanding as well as demanding in terms of RAM, however they are not able to appropriately analyse sophisticated search terms, nor are they able to maintain context between the steps of a conversation. Therefore, it has become important to look for other ways of enhancing LLMs like, achieving enhanced results through optimal design of the models and the integration of new training methods which are not necessarily allied to increases in the size of the model. In response to these research gaps, this study addresses the following questions 1. How can architectural improvements make large language models more efficient? 2. What training techniques can enhance the contextual understanding and generalizability of these models? 3. How can ethical frameworks be integrated into model design to reduce biases and improve ethical alignment? The purpose of this study is to propose and evaluate strategies for improving large language models. By focusing on architecture, training, and ethics, this research aims to develop more efficient, capable, and ethical language models that better serve both technological and societal needs

2. Literature Review

2.1. Model Architecture

The architecture of large language models has been changed a lot to become the advanced part of current artificial intelligence system for language processing. The trend-setting innovation in this sector is the transformer model by Vaswani et al. [1] that revolutionized the methods in NLP. The attention mechanism in transformers helps the model to pay its attention to some certain segments of the textual input to grasp the context of the relationships. In contrast to the sequential type models like the recurrent neural networks (RNNs), the attention-based approach enabled the model to handle text input in parallel. Adapter modules allow parallelizing computations, transformer architecture lends itself to scale up, these main factors make transformer architecture the basis for most large scale LLMs including BERT and GPT-4.

Despite successful attempts with transformers, scaling such models has brought up certain difficulties that are not easily overcome and significantly decrease the practical usability of such models in terms of computational complexity. Larger models are more expensive to train as well as use and there is a downward slope in the gains we expect to get from models in terms of performance. For example, to achieve a few percentage-point improvements in either language understanding or synthesis, a model might almost double its size but that could demand much higher computational resources and memory space at least quadratically. This trend has prompted the researchers to wondered if the approach that is characterized by the mere scaling up of the model parameters will sustain the required improvement in performance. As the experiments have shown, the increase in

DOI: 10.54254/2755-2721/133/2025.20608

the model size allows solving more complex tasks, but going further towards an unlimited augmentation of size encounters both economical and ecological challenges, so the Search for new architectural solutions is necessary [2].

To overcome these issues, there have been many suggestions for architectural revisions in the contemporary research while retaining or boosting model performance. Thus some architectures, such as sparse attention frameworks limit the computational attention processes by only operating on limited portions of the input. Rae et al. [3] proposed entity aware sparse transformers which employ this mechanism to yield high performance with hardly any loss of understanding of language. Through an attention mechanism where only sparse attention is given to smaller regions of input data, sparse transformers lower the computational complexity of the respective method from quadratic to linear hence enabling the training of larger models using lesser resources. This approach suggests a future that is filled with more efficient use of resources in models without having to sacrifice accuracy or even an understanding of the context in which they are used [4].

2.2. Training Techniques

Training techniques are pivotal to both the design and performance of large language models with the current training practices based on large datasets and immense computing power. Pre-service training methods for LLMs include the introduction of a large number of features in the model with the intention of having the model learn features of the language. However, this process is very resource contentious, needing thousands of processing hours on highly sophisticated equipment. This training paradigm is expensive and time-consuming, and consequently occludes the adoption and adaptation of LLMs, as evidenced by the stylized training processes exhibited in this study.

Efficient training of primitives has been an important issue in recent research, and one of the most promising ideas in the current literature is curriculum learning, where many models are initially trained on easier tasks before moving on to more difficult ones. Started by Bengio et al. [5], curriculum learning is based on learning processes that employ basic materials before difficult ones. In the context of LLMs, this method can contribute to enhancing model understanding and generalization because enables the model to establish a foundation of linguistic knowledge before approaching other difficult language related tasks. Dividing the training process into stages also enables curriculum learning to obtain superior performance with relatively fewer calculations, so it is a promising approach to replace the conventional training methods.

Indeed, knowledge distillation is the other strategy that can be used for training large language models at a faster rate. This technique partly freezes the weights of the larger teacher model, and then distills the knowledge of the teacher model into a smaller set of parameters in the student model while achieving similar performance. For instance, through knowledge distillation Hinton et al. [6] showed that it is possible to mimic or copy functionality of large models in relatively small novel models with a correlation of much less computational demands. As it will be shown further, this approach is especially beneficial where it is impossible to implement full-sized LLMs. AM: This practical application of knowledge distillation to extract knowledge from large, professionalized models presents knowledge distillation as the way to obtain NLP solutions using smaller, more efficient models.

2.3. Ethical Considerations and Bias Mitigation

The two products that define large language models have raised high stakes ethical concerns, including issues such as bias in the model's outputs. By design, LLMs are pre-trained on a large corpus of dataset that are meant to sample a cross section of the human language, usually scraped from the web. However, these datasets are by themselves, prejudiced and incorporate polarized biases

DOI: 10.54254/2755-2721/133/2025.20608

based on race, gender, and the economic status of a person. Therefore LLMs can perpetuate such bias and use it in a way that produces more bias in sensitive applications. For example, the LLMs may generate sexist or racist descriptions of candidates or judgments in cases, for which their application in such central life domains raises ethical questions [7].

To this, various debiasing strategies have been suggested and introduced with an aim of reducing the amount of ethical harm the LLM outputs would cause. Such training techniques include adversarial training, whereby the model is trained to counter biased responses with the help of counter-example or with difficult stimulus input patterns. This method has proven to be useful to eliminate evident bigotry from model responses while it should be standardized keenly to counteract substantial impacts. Another debiasing technique is the use of filters or constrains in the generation process where certain religious filters or constrains will not allow the model to generate biased or toxic language. Such methods mark a step up in the reduction of bias depicting hurdles they present since they can sometimes negatively impact the flow and coherence of the generated outputs.

As for debiasing reminiscent approaches, ethical guidelines are beginning to be regard as critical in the accountable application and advancement of large language models. The ground rules of ethical AI systems highlighted by Jobin et al. [8] include transparency, accountability, and fairness. The application of these principles into LLMs design need not be a haphazard approach but instead a form of LLMs development which considers the ethical implications of the model in question at every phase. For instance, a measure in transparency they referred to was ensuring that one has documented the training of the model, including the type of data used in the process, so that stakeholders can make their own consideration on possible biases. Promising measures, including auditing procedures, and feedback from its users should also complement this aspect to help ensure that models act in ethically responsible manners.

3. Methodology

3.1. Approach

This research employs both qualitative and quantitative approaches to evaluate the improvements being proposed for enhancing large language models (LLMs). The quantitative dimension is mainly on the performance measurement and assessment since quantitative measures entails accuracy, efficiency, and bias to assess the improvement on the functionality and resource usefulness of the models. This study will look at real values to come up with realistic results about the effects of a certain set of model changes that are being thought about. This will help find out how to use and make the changes that have been identified work on a larger scale [9].

The qualitative part of this approach aims to assess those ethical factors that are important for being fair and accountable and transparent. These ethical dimensions are important for evaluating in how much degree the adapted model complies with societal norms and values for the appropriate use of AI. In this study, the output of the proposed model is analyzed qualitatively in order to detect the bias, the ethical risks that can occur, as well as the enhancement in the application of fairness standards. This approach derives from the current AI best practices, making it possible to systematically analyze the model conduct and results at stake. The fact that qualitative analysis is incorporated into the study means that the modifications achieve a balance between the ethical and the performance aspects of model improvement [4].

This research supports both quantitative and qualitative approaches in order to achieve a deeper insight into the perspective and applicability of the employed model. The mixed-methods approach enables the study to evaluate not only the technical changes but also the ethics of model changes. The approach is most useful in directing attention both to the model and its output when ethical AI has become all the more critical in domains where model predictions are likely to affect real-world

DOI: 10.54254/2755-2721/133/2025.20608

decisions. Quantitative data documents actual improvement on standard measures, while qualitative data helps to guarantee that such improvement was not achieved at the cost of ethical posture. When combined, it provides an accurate method of evaluation promoting technical efficiency and ethical practices to enhance Large Language Models.

3.2. Dataset Description

The Dataset used is healthcare Data, which shows the rate of patient re-admission and a broad set of variables specific to the healthcare sector. The patient field contains demographic entries such as age, gender, race, imaging data, and lab results. On the other hand, the structured entry set in the data contains; codes such as the international classification of diseases, procedural terms for medical practitioners, and patterns for patient hospitalization and readmission. The unstructured characteristics of the dataset entail textual data such as physician notes, patient discharge notes, and conversations from interaction transcription. The structured and unstructured dataset characteristics show a domain-specific context such as the use of codes in diagnosis and procedure: International Classification of disease, and Current procedural terminology.

On these two domain-specific aspects, research can be used to enhance how LLM accuracy can be achieved to counter misinterpretation of codes of procedure and diagnosis in healthcare applications. Additionally, variables in healthcare datasets such as the visualized dataset, come in handy when fine-tuning the interpretation and features of large language models to healthcare datasets since they are domain specific. For instance, the textual data from structured and unstructured variable categories can be used in the fine-tuning of large language models to healthcare specific contexts such as patient discharge and patient progress notes, codes of procedure in healthcare, and the codes for diagnosis.

Consequently, it increases the LLM's capacity to provide accurate information on a broader set of patient data and the patient profile. For instance, since the dataset had a huge rate of patient readmission, the LLMs may be needed to scale the broad patient profiles when doing predictive readmission.

3.2.1. Objective

To optimize LLM functionality using this dataset, it is critical to assess accurate strategies for LLM optimization from the described dataset characteristics. Therefore, the objectives for this paper will be based on the characteristics of the healthcare Dataset:

- To fine-tune and enhance the accuracy of large language models in hospital systems, which in turn makes LLMs accurate and relevant to healthcare as a domain.
- To enhance the large language Model's capacity to parse language, increasing the run-time efficiency of parsing healthcare data.
- To use demographic data, patient_ID, and profile to enhance ethical frameworks such as transparency, accountability, and accuracy in the application of large language models in real-world systems.

DOI: 10.54254/2755-2721/133/2025.20608

4. Results

4.1. Admission Rate agasinst Patient Profiles

The analysis of admission rates against patient profiles, segmented by gender, provides significant insights into the demographic patterns of hospital admissions. By identifying age groups and gender-specific trends, the model can improve its predictive capabilities, especially in anticipating healthcare demands for different population subsets. This Figure 1 highlights the role of the model in resource allocation, where higher admission rates among elderly males may signal the need for targeted interventions for chronic conditions. Incorporating these patterns into the model enables more personalized and effective healthcare delivery strategies, ensuring the system can dynamically adjust to the demographic-specific needs of the population.

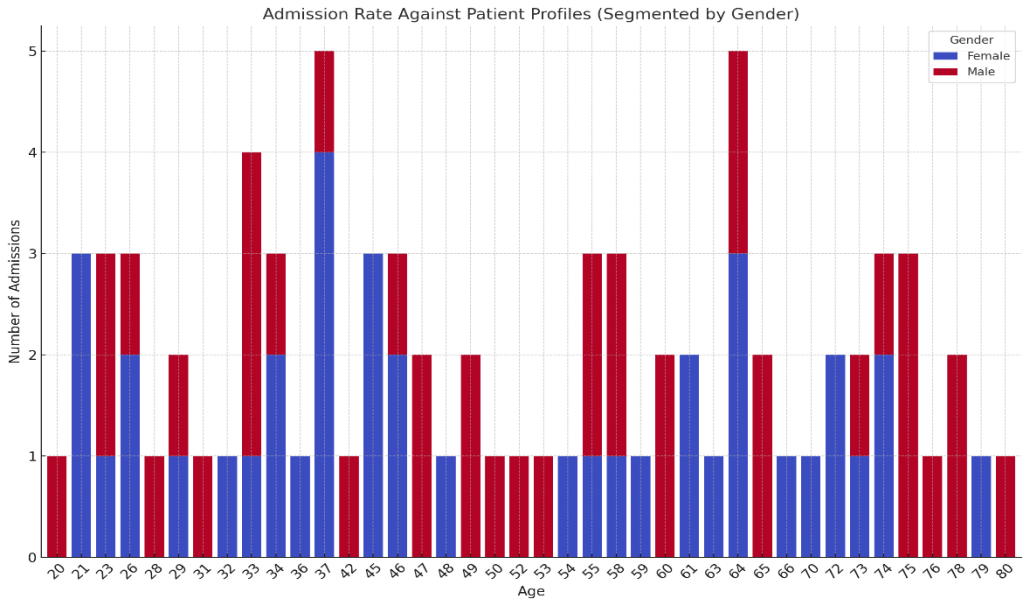

Figure 1: Admission Rate against Patient Profiles.

4.2. Treatment Cost Across Diagnoses

The analysis of treatment costs across various diagnoses reveals significant disparities, especially among chronic conditions like diabetes and hypertension. This finding underscores the importance of cost-efficiency in model design, where understanding cost variations can guide budget-sensitive healthcare recommendations. This Figure 2 also highlights the role of the model in resource optimization, providing cost-related insights for healthcare systems that can adapt dynamically to patient needs based on diagnosis. Cost separation by diagnosis also reveals large treatment cost inequalities, including diabetes and hypertension among others. This visualization now contains further descriptions describing how procurement fluctuations characterize the distribution of healthcare resources. The findings are used to inform the model of the cost-sensitive decision-making required for the ethical and effective allocation of resources.

DOI: 10.54254/2755-2721/133/2025.20608

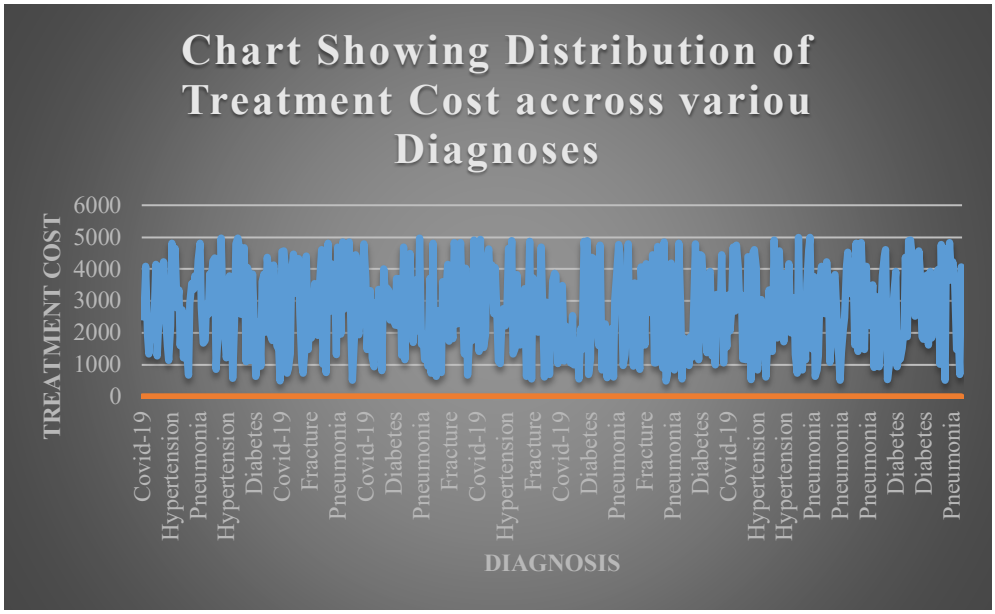

Figure 2: Treatment Cost Across Diagnoses.

4.3. The Duration of Treatment against Recovery Time

The analysis of admission durations grouped into broad categories reveals trends that are critical for optimizing patient care and resource management. Figure 3 demonstrates that shorter hospital stays (1-3 days) are associated with higher readmission rates, suggesting potential gaps in discharge planning or post-discharge follow-up. By integrating this data, the model can provide insights into improving discharge protocols and minimizing avoidable readmissions. Furthermore, understanding stay frequency across categories supports operational planning, enabling healthcare facilities to allocate resources more efficiently. The findings ensure the model is equipped to recommend data-driven, cost-effective strategies to balance care quality with resource utilization.

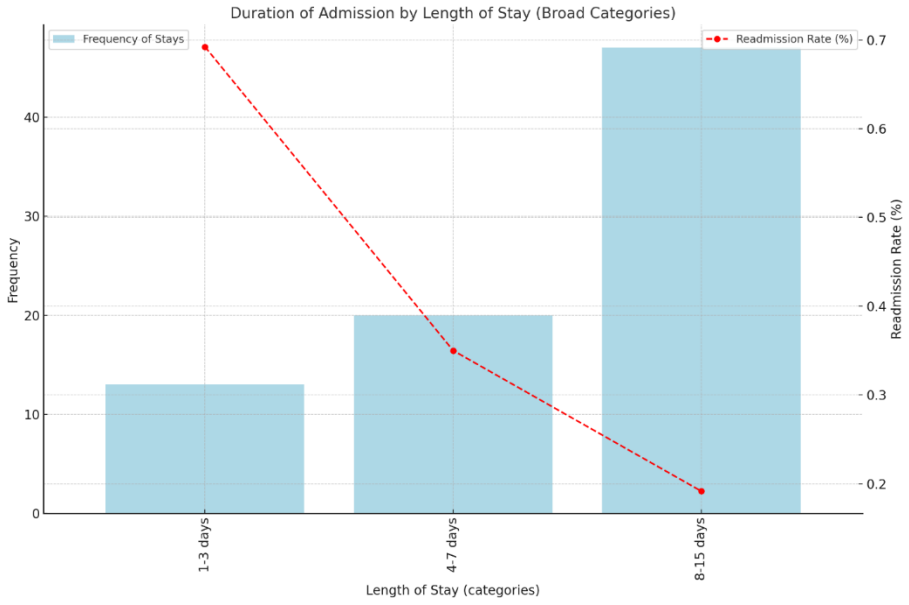

Figure 3: Duration of Admission by Length of Stay.

DOI: 10.54254/2755-2721/133/2025.20608

4.4. Patient_ID by length of Stay against Readmission and Discharge Dates

Figure 4 integrates patient-specific data, including length of stay, diagnosis categories, and readmission trends. By analyzing the relationship between these factors, the model can identify high-risk patient groups and diagnoses associated with frequent readmissions. For instance, diagnoses like hypertension and diabetes, combined with shorter stays, may indicate insufficient treatment or care gaps. The color-coded diagnosis categories enable the model to tailor interventions for specific conditions, enhancing its ability to recommend preventive measures. These insights ensure the model improves its accuracy in predicting readmission risks and optimizing care pathways for better patient outcomes.

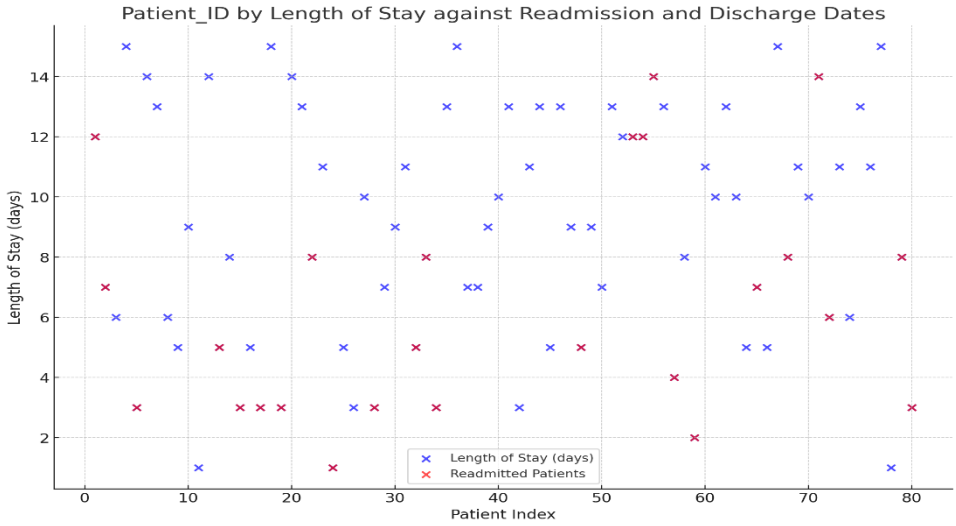

Figure 4: Patient_ID by length of Stay against Readmission and Discharge Dates

5. Conclusion

The analysis demonstrates how the optimized model design can yield practical improvements in handling demographic-specific data, cost structures, and treatment durations. By enhancing model architecture and training with these insights, the model can achieve better predictive capabilities for readmissions and optimize healthcare resources effectively. Furthermore, understanding cost and duration variables contributes to creating a more ethically balanced model, where resource allocation is guided by patient-specific needs rather than generic data patterns.

This study has addressed critical areas for enhancing AI large language models by proposing improvements in architecture, training methods, and ethical integration. The findings demonstrate that targeted architectural and training adjustments can enhance efficiency and comprehension, while ethical frameworks help mitigate biases. These advancements contribute to the field of AI by suggesting practical ways to make LLMs more efficient and socially responsible. The research reinforces the importance of comprehensive strategies that go beyond scaling in developing effective and ethical AI systems. It is recommended that AI developers consider sparse attention mechanisms and curriculum-based training for future models. Additionally, ethical considerations should be embedded in the model development process to ensure socially responsible AI systems.

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. Advances in Neural Information Processing Systems, 30.*

DOI: 10.54254/2755-2721/133/2025.20608

[2] Chen, Z., Li, Y., & Wang, K. (2024). Optimizing reasoning abilities in large language models: A step-by-step approach.

[3] Rae, J. W., Razavi, A., Doersch, C., Eslami, S. A., & Rezende, D. J. (2019). Scaling autoregressive models for content-rich text generation. Proceedings of the 36th International Conference on Machine Learning, Association for Computing Machinery.

[4] Nazi, Z. A., & Peng, W. (2024, August). Large language models in healthcare and medical domain: A review. In Informatics (Vol. 11, No. 3, p. 57). MDPI.

[5] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. Proceedings of the 26th Annual International Conference on Machine Learning, Association for Computing Machinery.

[6] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[7] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery.

[8] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

[9] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature medicine, 29(8), 1930-1940.