RESEARCH ARTICLE • OPEN ACCESS

Reconstruction of the Phi-2 Method for Question-Answering Related to Diabetes Disease Using the MedAlpaca Dataset

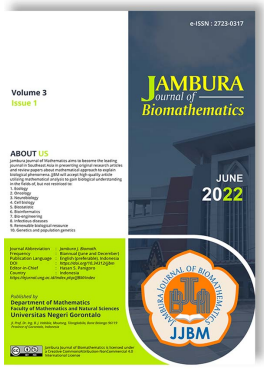
Muhammad Ridho, Alhadi Bustamam, and Risman Adnan

Volume 6, Issue 3, Pages 183–187, September 2025

Received 5 February 2025, Revised 18 March 2025, Accepted 23 June 2025, Published Online 1 September 2025

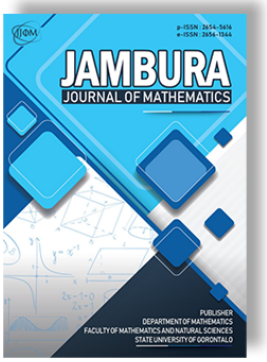
To Cite this Article : M. Ridho, A. Bustamam, and R. Adnan,"Reconstruction of the Phi-2 Method for Question-Answering Related to Diabetes Disease Using the MedAlpaca Dataset", *Jambura J. Biomath*, vol. 6, no. 3, pp. 183–187, 2025, *https://doi.org/10.37905/jjbm.v6i3.30506*

JOURNAL INFO • JAMBURA JOURNAL OF BIOMATHEMATICS



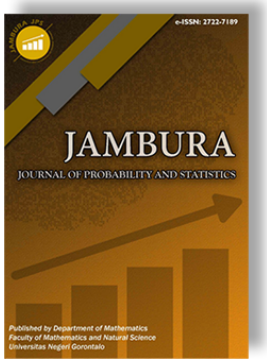
Homepage	:	http://ejurnal.ung.ac.id/index.php/JJBM/index
Journal Abbreviation	:	Jambura J. Biomath.
Frequency	:	Quarterly (March, June, September and December)
Publication Language	:	English
DOI	:	https://doi.org/10.37905/jjbm
Online ISSN	:	2723-0317
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjbm/oai
Google Scholar ID	:	XzYgeKQAAAAJ
Email	:	editorial.jjbm@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

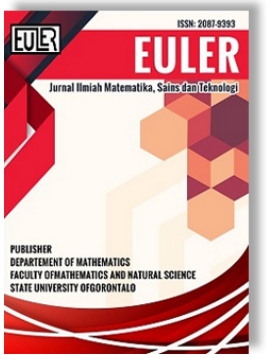


Jambura Journal of Mathematics

Jambura Journal of Mathematics Education



Jambura Journal of Probability and Statistics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi

Jambura Journal of Biomathematics, Volume 6, Issue 3, Pages 183–187, September 2025 https://doi.org/10.37905/jjbm.v6i3.30506

Research Article

Check for updates

Reconstruction of the Phi-2 Method for Question-Answering Related to Diabetes Disease Using the MedAlpaca Dataset

Muhammad Ridho[1,*], **Alhadi Bustamam**[1], **and Risman Adnan**[1,2]

[1]*Faculty of Mathematics and Natural Sciences, University of Indonesia, Depok, Indonesia*
[2]*Kalbe Digital Lab, PT Kalbe Farma, Tbk, Jawa Barat, Indonesia*

ARTICLE HISTORY
Received 5 February 2025
Revised 18 March 2025
Accepted 23 June 2025
Published 1 September 2025

KEYWORDS
Fine-Tuning
Phi-2
MedAlpaca
Question-Answering
Diabetes

ABSTRACT. *This study focuses on the reconstruction of the Phi-2 method for text-based question-answering systems related to diabetes using the MedAlpaca dataset. The aim is to enhance the accuracy in diabetes question-answering applications. We leverage LoRA techniques to fine-tune the model, thereby improving its ability to handle complex medical queries. The integration of the MedAlpaca dataset, which contains a diverse range of medical questions and answers, provides a robust foundation for training and testing the model. The results reveal that fine-tuning with MedAlpaca significantly enhances the model's performance, achieving higher accuracy compared to the base Phi-2 model, achieving a performance increase from 14.81% to 49.37% on MedMCQA, reaching 92.83% on PubMedQA, and 38.78% on MedQA. It also surpasses other leading models such as BioBERT (89.90%) and GatorTron (90.87%). The results highlight the effectiveness of incorporating domain-specific datasets like MedAlpaca to boost model performance. This advancement points to promising directions for future research, including expanding datasets and refining fine-tuning techniques to further improve automated medical question-answering systems.*

1. Introduction

In recent years, large language models (LLMs) have revolutionized the field of natural language processing (NLP) [1]. These models, which include millions or even billions of parameters, are trained on vast amounts of text data and have demonstrated remarkable capabilities in understanding and generating human-like text [2]. These advancements have opened up new possibilities for applications in various domains, including medical domain [3].

The application of LLMs in the medical domain holds immense potential for transforming healthcare delivery and medical research [4]. LLMs can assist in various tasks, such as extracting relevant information from medical literature, supporting clinical decision-making, and providing accurate and timely responses to patient queries, especially in deadly disease like diabetes [5].

Diabetes is a chronic disease that affects millions of individuals worldwide, posing significant challenges in terms of management and treatment [6]. The prevalence of diabetes is increasing significantly every year [7]. The complexity of diabetes care, which involves continuous monitoring, lifestyle adjustments, and medication adherence, necessitates the availability of accurate and reliable information for both patients and healthcare providers [8]. Given the prevalence and impact of diabetes, developing a specialized question-answering system for diabetes-related queries is of paramount importance [9].

On the other hand, Phi-2, a 2.7 billion-parameter language model that demonstrates outstanding reasoning and language understanding capabilities, showcasing state-of-the-art performance among base language models with less than 13 billion parameters [10]. With its compact size, Phi-2 is an ideal playground for fine-tuning experimentation on a variety of tasks, including in question-answering about disease like diabetes.

In this research, we fine-tune the Phi-2 model using the MedAlpaca dataset, specifically designed for medical applications, to develop a specialized question-answering system for diabetes-related queries. By tailoring the Phi-2 model to the nuances of diabetes-related information, we aim to enhance its ability to provide precise and contextually relevant answers to users' questions.

This research makes several key contributions. First, it demonstrates the significant impact of fine-tuning the Phi-2 model with the MedAlpaca dataset, achieving performance improvements. These enhancements highlight the model's effectiveness in a specialized medical domain and its versatility beyond general-purpose applications. Second, the study emphasizes the value of domain-specific datasets like MedAlpaca in enhancing the performance of large language models (LLMs) for targeted medical tasks. Third, it addresses a critical gap in medical question-answering systems by focusing on diabetes, a prevalent and complex chronic disease. By providing reliable and accurate information for diabetes-related inquiries, this system has the potential to support patients, healthcare providers, and researchers in making informed decisions and improving health outcomes.

*Corresponding Author.

Email : *mridho@sci.ui.ac.id* (M. Ridho)
Homepage : http://ejurnal.ung.ac.id/index.php/JJBM/index / E-ISSN : 2723-0317

2. Model And Datasets

2.1. Base Model: Phi-2

The Phi-2 is a Transformer-based model trained on 1.4 trillion tokens across multiple passes through a combination of synthetic and web datasets for natural language processing and coding tasks. With 2.7 billion-parameter language model that excels in reasoning and language comprehension, It achieves top-tier performance among base models with fewer than 13 billion parameters. Phi-2's performance on complex benchmarks rivals or surpasses that of models up to 25 times its size, thanks to innovative advancements in scaling and data curation. Its compact design makes Phi-2 an excellent tool for researchers, offering opportunities to explore mechanistic interpretability, safety enhancements, and fine-tuning across various tasks [10]. That's why we are interested in fine-tuning this model to medical question-answering related to diabetes.

In this work, we utilized the Phi-2 model developed by Microsoft, which can be accessed through the official repository. The model is licensed under the MIT License, a permissive license that allows for broad use and distribution, ensuring flexibility in research applications. Acknowledging these licensing terms ensures compliance with legal and ethical standards for the use of software in academic research.

2.2. Datasets: MedAlpaca

The is a curated collection of medical texts designed to support NLP applications in the healthcare domain. It includes a wide range of medical documents, research papers, clinical notes, and patient education materials [11]. This diversity ensures that the dataset provides a comprehensive overview of medical knowledge, making it an ideal choice for fine-tuning models aimed at improving performance in diabetes-related question-answering. The dataset consist of some medical data described below.

1. Medical Meadow Medical Meadow is a compilation of medical tasks designed for fine-tuning and assessing the performance of LLMs within the medical field. It is divided into two primary categories: a set of well-established medical NLP tasks that have been restructured into instruction-tuning formats, and a collection of various internet-sourced materials. Each dataset targets different facets of medical knowledge and practice, offering a thorough framework for both training and evaluation.
2. Dataset 1: Flash Cards Used by Medical Students Medicine involves various subjects that medical students must learn, including basic sciences, clinical knowledge, and skills. The Anki Medical Curriculum flashcards, made by medical students, cover the full medical school curriculum with topics like anatomy and pharmacology, using summaries and mnemonics to aid learning. These flashcards is used to create question-answer pairs for training by using GPT-3.5-Turbo to convert the remaining flashcards into relevant Q/A pairs in total 33,955 samples. These pairs are concise due to the flashcards' limited space.
3. Dataset 2: Stackexchange Medical Sciences The stack exchange dataset consists of 52,475 question-answer pairs obtained from five Stack Exchange forums related to biomedical sciences and related fields:
   Academia: This forum delves into research methods, the process of scientific publishing, and career trajectories within the scientific field. Although not solely focused on medicine, the extensive amount of medical research suggests that medical professionals will find models in this area useful.
   Bioinformatics: As a field that integrates biology, computer science, and data analytics, the Bioinformatics forum provides essential insights into techniques and tools for analyzing complex biological data, which is becoming crucial in contemporary medical research.
   Biology: Covering topics like genetics, physiology, and molecular biology, this forum contributes essential concepts to basic medical research. Including this forum helps integrate core life sciences into the training dataset.
   Fitness: This forum covers practical advice on maintaining and improving physical health through exercise routines, nutrition, and injury prevention. By including the Fitness forum, we expose models to health-related information that can be directly applied to patient care and lifestyle advice.
   Health: The Health forum encompasses a wide array of topics related to personal health, disease prevention, and medical treatments, offering information that can be directly utilized in medical practice.
4. Dataset 3: Wikidoc
   WikiDoc is a collaborative platform where medical professionals share and update medical knowledge. The platform has two main sections: the "Living Textbook" and "Patient Information." The "Living Textbook" features chapters on various medical specialties then using GPT-3.5-Turbo to convert the paragraph headings into questions and used the paragraphs as answers. In contrast, the "Patient Information" section already has questions as subheadings, so no rephrasing was needed. This dataset contains 10,000 samples.
5. Dataset 4: Medical NLP Benchmarks
   This includes: The COVID-19 Open Research Dataset Challenge (CORD-19), consisting of more than 800,000 scholarly articles, Benchmark data from Measuring Massive Multitask Language Understanding containing 3,787 samples [12], Training data from the Pubmed Causal Benchmark containing 2,446 samples [13], Conversational data from medical forums as presented in Chatdoctor [14], and The OpenAssistant dataset. A Crowd sourced conversational dataset, especially targeted towards training models with RLHF.

3. Data Preprocessing

The dataset was carefully preprocessed to ensure high-quality input for model training. Initially, the raw data was filtered to retain only the relevant medical topics, focusing on diabetes-related questions to align with the scope of the research. Following this, the data was cleaned by removing duplicates, resolving inconsistencies, and handling missing values. Tokenization and normalization were performed to standardize the input format, allowing seamless integration with the model architecture.

In addition to removing duplicates and handling missing values, several further steps were taken to ensure the quality of the MedAlpaca dataset during data cleaning. First, text normal-

ization was performed to standardize medical terminology and ensure consistency in the representation of abbreviations and medical terms across the dataset. This step was crucial for aligning the format of the questions and answers, which often vary in clinical texts. Next, stopword removal was applied to eliminate common words that do not contribute significant meaning to the model's understanding, such as "and," "the," or "of." This helped reduce noise in the dataset and streamline the input for more effective processing. Lastly, spelling correction was conducted to address any typographical errors or misspelled terms, which are often found in medical datasets. By correcting these errors, the data was made more reliable for training the model. These steps collectively enhanced the dataset's quality, ensuring a more accurate fine-tuning process.

After all, the final dataset was refined to 27,795 samples related to diabetes. These samples was then used as the training data for fine-tuning the model, ensuring that the processed data was both relevant and suitable for enhancing the model's performance on medical question-answering tasks related to diabetes.

4. LoRA: Low Rank Adaptation

To adapt pre-trained language models for our specific tasks, we used a technique called LoRA for updating the model's weights. LoRA keeps the original model weights frozen and adds special matrices to each layer. These matrices are much smaller than the original weights, requiring less memory and training time. This method is much more efficient than traditional fine-tuning, where all the weights are adjusted [15].

5. Fine-Tuning Process

The fine-tuning process for the Phi-2 model involves several key steps to adapt the pre-trained model for the specific task of diabetes-related question-answering. Initially, the pre-trained Phi-2 model is initialized with its existing weights and parameters. This model has been previously trained on a diverse range of texts, allowing it to develop a broad understanding of language.

Next, the model is fine-tuned on the diabetes-specific subset of the MedAlpaca dataset. During this phase, the model undergoes supervised learning, where it is trained to predict the correct answers to diabetes-related questions based on the context provided by the dataset. This involves adjusting the model's weights and parameters to minimize errors in its predictions, thereby enhancing its ability to generate accurate and relevant responses.

To ensure that the fine-tuning process is effective and the model generalizes well to new, unseen data, a separate validation set is used to evaluate the model's performance. This validation step helps in assessing whether the model is overfitting to the training data or if it can maintain its performance across different data points.

Finally, hyperparameters such as learning rate and batch size are optimized to achieve the best possible performance. To address this, adaptive learning rate methods like Adam is used, which adjust the learning rate based on gradient updates, thus enhancing training stability [16]. Next, gradient accumulation can be used to simulate larger batch sizes without increasing memory demands [17] and techniques such as cross-validation and monitoring validation performance are employed to select the optimal number of epochs and prevent overfitting [18]. Therefore, we got learning rate, 6 batch sizes, and 5 epochs. This optimization process fine-tunes the training dynamics to balance the model's learning speed and stability, ensuring that it performs optimally in answering diabetes-related queries.

By carefully executing these steps, the fine-tuning process aims to create a robust and reliable question-answering system that effectively supports diabetes-related information needs.

6. Evaluation and Comparison

For the evaluation of our fine-tuned model, we employed three prominent medical question-answering datasets: MedMCQA, PubMedQA, and MedQA. These datasets are widely recognized for testing the performance of models in the healthcare and biomedical domains, providing a robust benchmark for assessing the accuracy and relevance of language models in medical contexts.

- MedMCQA: a large-scale Multiple-Choice Question Answering (MCQA) dataset developed to tackle real-world medical entrance exam questions. Each question includes correct answer(s) and three distractors, challenging models to demonstrate deeper language comprehension [19]. In this study, we applied a filtering process to isolate questions specifically related to diabetes, yielding a subset of 2,734 questions along with their corresponding answer options.
- PubMedQA: an innovative biomedical question-answering (QA) dataset derived from PubMed abstracts. Each instance consists of (1) a question, which is either the title of an existing research article or based on one, (2) a context drawn from the abstract, excluding its conclusion, (3) a long answer in the form of the abstract's conclusion that addresses the research question, and (4) a yes/no/maybe summary answer [20]. In this research, we conducted a filtering process to extract questions focused exclusively on diabetes, resulting in a dataset of 4,298 questions along with their respective answer choices.
- MedQA: MedQA consists of multiple-choice questions similar to those found in medical exams, with four answer options provided for each question. It tests the model's ability to handle typical exam-style questions, which require precise medical knowledge and reasoning. The sources come from professional medical board exams [21]. In this study, we filtered the data to identify questions specifically related to diabetes, producing a dataset of 91 questions along with their corresponding answer options.

For comparison, we compared the performance of our fine-tuned model with several well-established biomedical models, including BioBERT [5], ClinicalBERT [22, 23], GatorTron [24], and BioLinkBERT [25]. Each model was evaluated using the MedMCQA, PubMedQA, and MedQA datasets.

7. Results

The results summarized in Table 1, highlight the significant improvements achieved by our fine-tuned model compared to the baseline Phi-2 model and the other models. The metrics used for comparison are accuracy percentages across the different datasets.

Table 1 presents the performance of various biomedical

Table 1. Accuracy comparison between baseline model Phi-2, fine-tuned model Phi-2 + MedAlpaca, and another biomedical models on MedMCQA, PubMedQA, and MedQA datasets. Best model in bold and second-best underlined.

	MedMCQA	PubMedQA	MedQA
Phi-2	14,81 %	69,29 %	19,39 %
Phi-2 + MedAlpaca	<u>49,37 %</u>	**92,83 %**	**38,78 %**
BioBERT	43,45 %	89,90 %	35,94 %
ClinicalBERT	32,58 %	84,03 %	34,90 %
BlueBERT	27,71 %	88,92 %	17,45 %
GatorTron	**51,35 %**	<u>90,87 %</u>	<u>36,84 %</u>
BioLinkBERT	28,63 %	86,00 %	29,41 %

models across three datasets. Notably, the Phi-2 model fine-tuned with the MedAlpaca dataset significantly outperforms the base Phi-2 model across all three datasets. For MedMCQA, the performance improves from 14.81% to 49.37%, demonstrating that fine-tuning with MedAlpaca greatly enhances the model's ability to answer diabetes multiple-choice questions. On PubMedQA, the Phi-2 + MedAlpaca achieves 92.83%, surpassing all other models, including BioBERT, which reached 89.90%, and GatorTron, at 90.87%. Similarly, for MedQA, the fine-tuned model achieves 38.78%, again a substantial improvement over the base Phi-2 score of 19.39%.

These results highlight the effectiveness of our fine-tuned Phi-2 + MedAlpaca model, which achieves state-of-the-art performance in several tasks, especially on the PubMEDQA and MEDQA dataset.

8. Future Works

These preliminary results are part of an ongoing study aimed at fully evaluating and optimizing the Phi-2 model for diabetes-related question-answering. Future works will involve:

- Completing the fine-tuning process with an expanded dataset, including abstracts of papers and additional medical knowledge data.
- Explore advanced fine-tuning methods like multi-task learning and domain adaptation.
- Using another metrics evaluations to measure the model's performance in quality and quantity.
- Further improving the model's robustness, user interaction, and integration with clinical systems.
- Develop integration strategies for embedding the model into healthcare systems like EHRs or CDSS.
- Collect feedback from healthcare professionals and patients for model refinement.

9. Conclusion

In summary, we have demonstrated the significant impact of fine-tuning the Phi-2 model with the MedAlpaca dataset on its performance across multiple biomedical question-answering tasks related to diabetes disease. Our results indicate that this fine-tuning approach leads to substantial improvements over the base Phi-2 model as well as other state-of-the-art models.

The enhancements observed in our study have important implications for biomedical applications. By significantly improving the accuracy and relevance of responses, our model holds promise for advancing automated medical question-answering systems, especially in the context of diabetes care. The ability to provide precise and contextually appropriate answers can greatly support medical professionals and improve patient outcomes.

Looking ahead, there are several avenues for future research. Expanding the dataset to include more medical texts and abstracts could further enhance the model's generalizability and performance across various medical topics. Additionally, refining fine-tuning techniques and experimenting with different hyperparameters might yield further improvements. Comparative analyses with other emerging biomedical models could also validate the effectiveness of our approach and identify opportunities for further advancements.

In conclusion, the results of this study provide compelling evidence of the benefits of fine-tuning the Phi-2 model with the MedAlpaca dataset. This approach not only improves performance across key biomedical question-answering tasks but also sets a new benchmark in the field, offering valuable insights for future research and application.

Author Contributions. **Ridho, M.**: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, project administration. **Bustamam, A.**: software, validation, supervision, funding acquisition. **Adnan, R.**: software, validation, supervision, funding acquisition.

Acknowledgement. The author would like to express sincere gratitude to Prof. Alhadi Bustamam, S.Si., M.Kom., Ph.D., Head of the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Indonesia, Head of the Data Science Center, University of Indonesia, and primary supervisor, for his invaluable knowledge, guidance, encouragement, and support that greatly contributed to the successful completion of this paper. Special thanks are also extended to Dr. Risman Adnan Mattotorang, S.Si., M.Si., Ph.D., as the second supervisor, for his expertise, insightful feedback, motivation, and continuous support throughout the preparation of this paper. Finally, the author would like to thank all members of the Data Science Center, University of Indonesia, and the Bioinformatics Laboratory, Department of Mathematics, University of Indonesia, for their assistance and support during the course of this research.

Funding. This research was funded by Data Science Center, University of Indonesia and the Bioinformatics Laboratory.

Conflict of interest. The authors declare no conflict of interest.

Data availability. Not applicable.

Abbreviations.

LLMs : Large Language Models.
LoRA : Low Rank Adaptation.

NLP : Natural Language Program.

References

[1] A. Vaswani *et al.*, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. DOI:10.48550/arXiv.1706.03762

[2] K. M. Fitria, "Information retrieval performance in text generation using knowledge from generative pre-trained transformer (gpt-3)," *Jambura Journal of Mathematics*, vol. 5, no. 2, pp. 327–338, 2023. DOI:10.34312/jjom.v5i2.20574

[3] U. Rifanti *et al.*, "A reinforcement learning based decision-support system for mitigate strategies during covid-19: A systematic review," *Jambura Journal of Biomathematics (JJBM)*, vol. 6, no. 1, pp. 60–70, 2025. DOI:10.37905/jjbm.v6i1.30513

[4] E. Alsentzer *et al.*, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019. DOI:10.48550/arXiv.1904.03323

[5] J. Lee *et al.*, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2019. DOI:10.1093/bioinformatics/btz682

[6] ADA, "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020," *Diabetes Care*, vol. 43, no. Supplement_1, pp. S14–S31, 2020. DOI:10.2337/dc20-S002

[7] S. Syarofina *et al.*, "The distance function approach on the minibatchk-means algorithm for the dpp-4 inhibitors on the discovery of type 2 diabetes drugs," *Procedia Computer Science*, vol. 179, pp. 127–134, 2021. DOI:10.1016/j.procs.2020.12.017

[8] M. J. Davies *et al.*, "Management of hyperglycemia in type 2 diabetes, 2018. a consensus report by the american diabetes association (ada) and the european association for the study of diabetes (easd)," *Diabetes Care*, vol. 41, no. 12, pp. 2669–2701, 2018. DOI:10.2337/dci18-0033

[9] IDF, *IDF Diabetes Atlas (10th ed).* Russels: International Diabetes Federation, 2021.

[10] Microsoft, "Phi-2: The surprising power of small language models," 2023, Accesed on 5 February 2025.

[11] Han *et al.*, "Medalpaca – an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023. DOI:10.48550/arXiv.2304.08247

[12] D. Hendrycks *et al.*, "Measuring massive multitask language understanding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. DOI:10.48550/arXiv.2009.03300

[13] B. Yu, Y. Li, and J. Wang, "Detecting causal language use in science findings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4663–4673 2019. DOI:10.18653/v1/D19-1473

[14] L. Yunxiang *et al.*, "Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge," *Cureus*, 2023. DOI:10.7759/cureus.40895

[15] E. J. Hu, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. DOI:10.48550/arXiv.2106.09685

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015. DOI:10.48550/arXiv.1412.6980

[17] J. Dean *et al.*, "Large scale distributed deep networks," in *Proceedings of the 26th Conference on Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012.

[18] Y. Bengio, "*Practical recommendations for gradient-based training of deep architectures*." In Neural Networks: Tricks of the Trade, vol. 7700, pp. 437–478, Heidelberg: Springer, 2012. DOI:10.1007/978-3-642-35289-8_26

[19] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," In *Proceedings of Machine Learning Research*, vol. 174, pp. 248–260, 2022.

[20] Q. Jin *et al.*, "Pubmedqa: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019. DOI:10.48550/arXiv.1909.06146

[21] D. Jin *et al.*, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021. DOI:10.3390/app11146421

[22] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019. DOI:10.48550/arXiv.1904.05342

[23] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets," In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 58–65, 2019. DOI:10.18653/v1/W19-5006

[24] X. Yang *et al.*, "Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records," *npj Digital Medicine*, vol. 5, no. 1, p. 194, 2022. DOI:10.1038/s41746-022-00742-2

[25] M. Yasunaga, J. Leskovec, and P. Liang, "Linkbert: Pretraining language models with document links," *arXiv preprint arXiv:2203.15827*, 2022. DOI:10.48550/arXiv.2203.15827