

UNIVERSITY OF ST.GALLEN

School of Computer Science

Improving Clinical Knowledge in Large Language Models through Incremental Learning Methods

Submitted by:

Ziwei Chen

22-601-660

Approved on Application by

Professor:

Prof. Dr. Christina Niklaus

Dr. Bernhard Bermeitinger

Date of Submission:

18/11/2024

Abstract

Table of Contents

[1 Introduction 0](#_Toc20896)

[1.1 Literature Review 0](#_Toc25838)

[1.2 Research Gaps 1](#_Toc4783)

[1.3 Research Questions 2](#_Toc32029)

[2 Datasets 3](#_Toc8884)

[3 LLMs Selection: BioMistral-7B 4](#_Toc888)

[4 Methodology 5](#_Toc11266)

[4.1 Prompt Engineering 5](#_Toc21397)

[4.2 Incremental Learning 6](#_Toc13366)

[5 Evaluation 8](#_Toc7029)

[6 References 8](#_Toc23745)

[7 Appendix 10](#_Toc15471)

# 1 Introduction

## **1.1** Literature Review

The advent of Large Language Models (LLMs) has marked a significant shift in the field of Natural Language Processing (NLP), particularly within the medical domain. These models, such as ChatGPT and Vicuna, have demonstrated remarkable versatility and advanced capabilities, exhibiting human-like comprehension and reasoning abilities that enable them to tackle a range of tasks from basic textual understanding to complex problem-solving challenges (OpenAI, 2023; Zheng et al. 2023). The transformative potential of open-source LLMs like LLaMA has facilitated their innovative use in specialized domains, including medicine (Workshop et al. 2023; Touvron et al. 2023a; Dave et al. 2023).

However, the integration of LLMs into healthcare presents unique challenges and opportunities. While preliminary adoption has opened new avenues for innovation, concerns about data privacy risks associated with proprietary models have arisen (He et al. 2023; Zhou et al. 2024). Initiatives like PMC-LLaMA and MedAlpaca have been developed in response to the community's interest in specialized LLMs for healthcare (Wu et al. 2023; Han et al. 2023). Yet, the adoption of open-source medical models has been limited, primarily due to the lack of lightweight models that can perform comparably to larger or proprietary models (Black et al. 2022; Touvron et al. 2023b; Jiang et al. 2023).

Building upon this foundation, the current work focuses on improving clinical knowledge in LLMs through incremental learning methods. This approach aims to enhance the models’ ability to understand and generate medically relevant content without compromising their general capabilities. BioMistral 7B, a more lightweight yet powerful specialized LLM tailored for the biomedical domain, serves as a pivotal starting point for this research. Derived from Mistral 7B Instruct v0.1 and further pre-trained on PubMed Central, BioMistral 7B not only demonstrates the potential for continuous improvement in clinical knowledge within LLMs but also addresses the need for more accessible, less resource-intensive models in healthcare settings (Jiang et al. 2023; Labrak et al., 2024). By focusing on a more compact architecture while maintaining high performance, BioMistral 7B offers a promising solution to the challenges of integrating LLMs into clinical practice.

The current study aims to leverage incremental learning methods to further refine the clinical knowledge base of BioMistral, ensuring that the model remains up-to-date with the latest medical advancements while being adaptable to various medical NLP tasks. This research is poised to contribute to the ongoing discourse on enhancing the capabilities of LLMs in the medical field through innovative training methodologies.

## **1.2** Research Gaps

The current project has the potential to fill a number of important research gaps in the field of medical artificial intelligence, which are necessary for the development of clinical expertise and the useful implementation of Large Language Models (LLMs) in healthcare environments. The foundation of this research project is the transformative potential of LLMs in comprehending clinical data, adjusting to the intricacy of clinical notes, and continually improving through incremental learning.

**Depth of Clinical Understanding:** A significant research gap exists in the depth of clinical understanding that LLMs can achieve. This pertains to the nuanced comprehension of clinical data, which includes the intricate details of disease symptoms, diagnostic procedures, and the efficacy of treatments. The complexity of medical language and the variability in symptom presentation across patients pose challenges that current models have yet to fully address (Labrak et al., 2024). Enhancing the models' ability to discern these subtleties is crucial for improving diagnostic accuracy and treatment plans.

**Adaptability to Clinical Notes:** Another gap that this project aims to address is the adaptability of LLMs to interpret unstructured clinical notes. Clinical notes are rich in detailed, narrative-style information that often diverges from the structured datasets traditionally used for model training. There is a pronounced need for LLMs to effectively navigate and glean knowledge from this diverse and complex data, which is essential for providing comprehensive patient care (Touvron et al., 2023b).

**Incremental Learning for Continuous Improvement:** The capacity for LLMs to incrementally learn from new data is a critical area that requires further exploration. In the rapidly evolving field of healthcare, where medical knowledge and best practices are continuously updated, LLMs must be able to adapt and refine their knowledge base accordingly (Jiang et al., 2023). This capability is vital for maintaining the relevance and reliability of LLMs in medical applications.

To address these gaps, the 'augmented-clinical-notes' datasets available on Hugging Face serves as a valuable resource for this project. This dataset, which encompasses a wide array of clinical notes, provides an in-depth perspective on patient symptoms, diagnoses, treatments, and outcomes. The breadth and depth of this datasets render it an excellent candidate for training and evaluating the performance of generative models within the medical domain (Hugging Face, n.d.).

By employing precise prompting engineering (PE), we aim to extract pivotal knowledge points from these clinical notes. Furthermore, we will convert the unstructured clinical notes into structured JSON format, thereby enabling a more organized and analyzable representation of the data. This knowledge will then be utilized to incrementally train BioMistral-7B, with the expectation of significantly bolstering the model's clinical knowledge base and its capacity to generate informative and structured medical summaries (Beltagy et al., 2020).

The effectiveness of the incremental training will be rigorously evaluated using the Supervised Fine-tuning Benchmark datasets mentioned in the BioMistral-7B paper. These datasets, accessible at bigbio/med\_qa and openlifescienceai/medmcqa, offer a comprehensive framework for benchmarking the performance of medical LLMs (Fries et al., 2022; Singhal et al., 2023a). They provide a robust set of medical QA tasks that will enable a thorough assessment of the improvements in BioMistral-7B's capabilities post-incremental training.

By bridging these research gaps, this project aspires to enhance the clinical knowledge understanding of BioMistral-7B, equipping it with the ability to process and summarize medical records more effectively. This enhancement is anticipated to contribute to more informed clinical decisions and, ultimately, improved patient outcomes.

## **1.3** Research Questions

The research objectives of this project are delineated by the following questions, which will steer the exploration and development process. These questions are designed to maintain a concentrated and purposeful methodology aimed at augmenting the capabilities of BioMistral-7B for clinical knowledge tasks:

**RQ1: How can incremental learning be effectively integrated into LLMs to improve their understanding of clinical narratives?**

This question aims to explore the feasibility and methods of implementing incremental learning within the BioMistral-7B model. The goal is to determine the best practices for continuously updating the model's knowledge base with new clinical data. Specifically, this incremental learning will utilize self-supervised training techniques, enabling the generative large model to better comprehend domain-specific knowledge before fine-tuning on downstream tasks.

**RQ2: What are the most effective prompt engineering strategies for extracting relevant medical information from unstructured clinical notes?**

This research question focuses on developing and refining prompt engineering techniques to maximize the extraction of key medical details from clinical notes. The aim is to identify prompts that lead to the most accurate and comprehensive data structuring. To achieve this, we will explore how core statements within clinical notes, supported by contextual background, can be effectively highlighted through prompt engineering. This approach will enable the large language model to not only extract data but also to better understand the causal relationships within the medical domain. By doing so, the model will be able to grasp the underlying mechanisms that connect symptoms, diagnoses, treatments, and outcomes, thereby enhancing its ability to process and summarize clinical narratives in a manner that is coherent with the domain's knowledge structure.

**RQ3: To what extent can a structured format of clinical notes enhance the model's ability to generalize and adapt to new, unseen medical data?**

This research question explores the impact of using a structured format, such as JSON, to enhance the BioMistral-7B model's capacity to generalize and adapt to novel medical data. The structured format is designed to capture the complexity and nuances of clinical narratives by articulating critical relationships and causal links among medical entities. This approach mirrors the associative capabilities of knowledge graphs, allowing the model to encapsulate core assertions within their contextual backdrop. The investigation will assess whether this method can provide the large language model with a more profound understanding of the interconnections within clinical data, similar to the effects achieved by knowledge graphs. By doing so, this research will offer insights into the potential of structured data representations to bolster clinical knowledge comprehension, especially in the absence of mature medical knowledge graph models.

**RQ4: How does the performance of the incremental pretrained medical LLMs compare to the original model on standardized medical question-answering tasks, and what are the potential limitations and ethical considerations of using such a model in real-world clinical settings?**

This research question is designed to assess the effectiveness of incremental training on a medical language model, specifically BioMistral-7B, by comparing its performance against the original model on standardized medical QA tasks. The primary objective is to quantify the improvements in the model's ability to understand clinical knowledge and to summarize medical records effectively. Additionally, this question aims to investigate the potential limitations and ethical considerations associated with the deployment of AI models in real-world clinical settings. By examining these factors, the research will provide a thorough evaluation of the benefits and challenges of using incrementally trained medical language models in practical healthcare environments, with a focus on ensuring their application is both effective and adheres to ethical standards.

# 2 Datasets

The datasets employed in this study is the 'augmented-clinical-notes' datasets, which is part of the Hugging Face datasets collection and can be referenced as AGBonnet/augmented-clinical-notes. This datasets comprises a substantial compilation of 30,000 authentic clinical notes, serving as an invaluable resource for the training and evaluation of generative models within the medical domain. The notes within this datasets exhibit a wide variety of medical conditions and treatments, making it an exemplary datasets for the development of a robust medical Large Language Model (LLM).

The lengths of the complete clinical notes in this datasets range from 746 to 31,000 words, with each note detailing symptoms, diagnostic findings, treatment methods, and outcomes. This datasets is not only highly valuable for its medical content but also aligns well with the objective of constructing structured inputs for model training, capturing the interrelationships between various medical entities. To illustrate the nature of the data, a summary of a case from the datasets is provided below:

**Case Summary**: Amidst a complex medical history of metastatic renal cell carcinoma, a 67-year-old patient presented with shortness of breath, pleuritic chest pain, and left scapular pain. Diagnostic findings revealed a gastro-pleural fistula between the stomach and pleural space, as well as multiple metastases and atelectasis. Treatment involved a novel approach utilizing a venting gastrostomy tube and chest tube to water seal, closure attempted with endoscopic suturing, followed by laparoscopic surgery for fistula repair. The patient's postoperative course was successful, with closure of the fistula, and they were discharged to a rehabilitation facility. Four months of follow-up included the patient tolerating an oral diet, with the removal of the gastrostomy tube, jejunostomy tube, and chest tube without complication.

This case illustration exemplifies the depth and breadth of the 'augmented-clinical-notes' datasets, highlighting its utility for training LLMs to understand and generate structured medical narratives that mirror the intricacies of real-world clinical practice. The comprehensive nature of the datasets positions it as an ideal resource for developing models capable of processing and summarizing clinical information in a manner that aligns with established medical knowledge structures.

While this datasets is highly suitable due to its rich medical content, it is important to note that the text notes are not structured content and cannot be directly fed into the model in their entirety. Doing so would be inefficient, as it would include much irrelevant content. Therefore, a more structured approach is needed, which is where prompting engineering (PE) with a general large model like ChatGPT comes into play. By using PE, we can annotate the entire datasets, transforming unstructured clinical notes into structured information that can be more effectively utilized for model training.

# 3 LLMs Selection: BioMistral-7B

In this study, we have chosen to employ the BioMistral-7B model, a state-of-the-art generative Large Language Model (LLM) developed by Labrak et al. (2024), known for its exceptional performance in processing complex biomedical and clinical text. Built upon the Mistral 7B Instruct v0.1 model, BioMistral-7B is designed for prompt instruction incorporation and fine-tuning across a variety of tasks. Its extensive pre-training on the PubMed Central corpus equips it with a comprehensive understanding of medical literature, making it an ideal candidate for our research in the medical domain.

The choice of BioMistral-7B as the base model is highly appropriate for several reasons. Firstly, its extensive pre-training on the PubMed Central corpus provides it with a deep understanding of medical literature, which is crucial for our research in the medical domain . Secondly, the model has demonstrated superior performance in processing biomedical and clinical text, which is a key requirement for our study . Additionally, BioMistral-7B has undergone rigorous evaluation on a benchmark consisting of 10 established medical question-answering (QA) tasks in English, outperforming existing open-source medical models . The model's capabilities extend beyond English, as it has been assessed in multiple languages, showcasing its robustness across diverse linguistic contexts .

To enhance the model's accessibility and practicality, BioMistral-7B incorporates lightweight models obtained through quantization and model merging approaches. These techniques are pivotal for deploying the model on consumer-grade devices, ensuring that the benefits of advanced language models can be realized in various real-world medical applications . The model has been reported to have a lightweight parameterization, which, combined with its medical data pre-training, makes it highly suitable for our needs . The use of quantization and model merging also ensures that the model remains efficient and effective, even when deployed on devices with limited computational resources .

In summary, BioMistral-7B's strong performance in biomedical text processing, its multilingual capabilities, and its lightweight design make it an excellent choice for our study. These features, along with its ability to be fine-tuned and incorporated with prompt instructions, position it as a powerful tool for generating informative and structured medical summaries.

# 4 Methodology

## 4.1 Prompt Engineering

The overarching aim of this study is to engineer a structured approach for the analysis of clinical notes, enhancing the granularity and relational clarity of patient data. Given that raw clinical notes are unstructured text, their direct use in incremental learning would be inefficient and could obscure critical information with irrelevant details. To address this, we propose a method for annotating and transforming these unstructured notes into a standardized JSON format. This structured format is intended to encapsulate the essence of each patient case by distilling detailed and specific information, thereby mitigating the inefficiencies associated with processing long, unformatted text.

Our method involves segmenting the data into distinct categories such as Chief Complaints, Medical History, Diagnostic Findings, Diagnosis, Treatment, and Outcome, with each category delineated by a defined set of sub-fields. This structured JSON format will serve as the exclusive input for our model, effectively rendering raw text data obsolete.

To achieve this transformation, we employ prompt engineering techniques, leveraging a foundational large language model to annotate the clinical notes. This process involves using the GPT-3.5 Turbo model to generate structured annotations, consuming approximately 100 million tokens in prompts and taking roughly 40 hours to produce the training data. The result is a datasets that is not only organized but also retains the critical relationships between various pathologies, treatments, and outcomes. By providing this structured input, we anticipate that the model's training and incremental learning processes will be significantly enhanced, leading to more accurate identification of patterns, correlations, and dependencies between different aspects of patient care. This, in turn, is expected to improve diagnostic and treatment predictions, as well as the model's generalization capabilities when encountering new, unseen data.

The adoption of this structured data format will streamline the data preprocessing stage and provide a robust foundation for building a model that can scale and adapt to the evolving complexities of clinical data management. An example of the structured JSON output is as follows:

{

"PatientInformation": {

"ChiefComplaints": [

"Complaints of pain and swelling in the right back for several weeks",

"No significant health problems except a thoracic trauma one year prior"

],

"MedicalHistory": {

"PreviousInjury": "Thoracic trauma with a simple fracture of the 9th right rib"

},

"DiagnosticFindings": [

{

"Test": "X-ray",

"Finding": "A shadow in the lower part of the right hemithorax"

},

{

"Test": "CT-scan",

"Finding": "A tumor with heterogeneous density and destruction of the 9th rib"

}

]

},

"Diagnosis": {

"Disease": {

"Name": "Sclerosing xanthofibroma",

"Type": "Benign tumor",

"Location": "Thoracic wall"

}

},

"TreatmentAndOutcome": {

"Treatment": {

"Type": "Surgical resection and plastic repair",

"Details": "Involving three ribs and reconstruction with polypropylene mesh"

},

"Postoperative Course": {

"Recovery": "Uneventful",

"DischargeStatus": "Good condition"

},

"FollowUp": {

"Duration": "Two years",

"FunctionalStatus": "Patient returned to work one month after surgery"

}

}}

This structured format exemplifies the depth and clarity of the data that our model will process, providing a clear framework for the model to learn from and generate insights, thereby enhancing its clinical knowledge and predictive capabilities.

## 4.2 Incremental Learning

In this segment of our methodology, we concentrate on the incremental learning process, which is facilitated by self-supervised pretraining on the structured JSON data. This self-supervised approach allows the model to implicitly learn from the transformed data without explicit labels, thereby acquiring new knowledge and enhancing its understanding of clinical information. Following this pretraining phase, we proceed to evaluate its performance on QA tasks through supervised fine-tuning.

**Data Preparation and Tokenization**

For the incremental learning phase outlined in this research, we utilize the structured JSON data obtained through Prompt Engineering (PE) techniques as described in the preceding sections. This JSON data, which encapsulates key medical details extracted from clinical notes, serves as the primary input for our model training.

A critical aspect of our data preparation involves determining the appropriate sequence length for training. After careful consideration and analysis, we have established that a maximum sequence length of 1024 tokens is adequate to encompass the entirety of the information within the JSON-formatted data. This determination is rooted in the observation that padding the post-PE JSON data to a length of 1024 tokens ensures that all pertinent data points are retained without any loss of information. This approach is feasible because the tokenizer can effectively process the structured JSON data, which, although derived from original medical notes potentially reaching up to 30,000 words, is condensed into a more focused and detailed JSON format after PE.

By adopting a sequence length of 1024 tokens, we ensure that our model can effectively handle the structured data without the need for excessive padding or truncation, which could otherwise introduce biases or omit crucial medical details. This sequence length strikes a balance between preserving the integrity of the medical information and maintaining computational efficiency during the training process.

To facilitate the subsequent self-supervised training of our model, we employ the tokenizer from the original model, BioMistral-7B. This tokenizer is utilized for the tokenization of the structured JSON data. By using the BioMistral-7B tokenizer, we ensure consistency in the tokenization process, which is essential for the model to understand and learn from the data effectively. The tokenizer's familiarity with medical terminology and context aids in the accurate representation of the structured data, thereby enhancing the model's ability to acquire new knowledge during the self-supervised training phase.

**Model Architecture and Training Strategy**

The model employed in this study is the BioMistral-7B, which boasts a 36-layer transformer architecture that is particularly adept at managing the intricacies of medical text. This deep neural network structure is chosen for its capacity to capture the nuanced patterns and relationships within the medical domain. To strike a balance between maintaining the model's core understanding and allowing for the assimilation of new information, we have elected to freeze the weights of the initial 24 layers. This strategy allows the foundational layers to retain their general understanding while the upper 12 layers are left unfrozen, enabling them to adjust and learn from the newly introduced structured JSON data. Due to the selective unfreezing of only 12 layers, the total number of trainable parameters amounts to approximately 2 billion, which are the parameters of the adjusted 12 layers. This focused training approach allows for a more efficient use of computational resources and faster convergence during the training process.

The training regimen is designed to span 5 epochs, which is deemed adequate for the model to thoroughly absorb the structured data and enhance its knowledge base. This iterative process is carried out on a single A800 80G GPU, a choice that is well-suited to meet the intensive computational requirements of training a model of this caliber. During the training, each batch consists of 16 samples, and the GPU is utilized at its full capacity, operating at 100% utilization. This ensures that the training is not only efficient but also maximizes the throughput of the GPU, leading to a more expedited and effective training session. The entire training duration up to 36 hours, providing the model with ample time to converge and develop a comprehensive grasp of the input data.

**Results and Model Selection:**

Throughout the training process, which comprised approximately 30,000 steps, the train loss consistently decreased, indicating a strong fitting effect as the model learned from the data. However, the valid loss exhibited a different trend: it initially decreased, reaching its nadir at around 11,000 steps, before subsequently increasing, suggesting the onset of overfitting. In response to this observation, we decided to use the checkpoint from 10,000 steps as the final model, as it represented the point of optimal performance on the validation set. This decision was made to avoid overfitting and to ensure that the model would generalize well to unseen data. The selected checkpoint thus serves as the trained model that balances learning from the new data with the preservation of the model’s original capabilities. Detailed records of the training and valid loss results are logged in wandb, and a summary is provided in the appendix for reference.

# 5 Evaluation

# 6 References

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, K., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Reynolds, L., Tow, J., Wang, B., & Weinbach, S. (2022). GPT-NeoX-20B: An open-source autoregressive language model. In Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models (pp. 95-136). Association for Computational Linguistics.

Dave, T., Anirudh, S. A., & Satyam, S. (2023). ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence, 6, 1169595.

Han, T., Adams, L. C., Papaioannou, J.-M., Oberhauser, T., Truhn, A., Löser, D., & Bressem, K. K. (2023). MedAlpaca – An open-source collection of medical conversational AI models and training data.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E. (2023). A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics.

Jiang, G., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv preprint arXiv:2402.10373.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023a). Capabilities of GPT-4 on medical challenge problems.

OpenAI. (2023). ChatGPT: Language models are few-shot learners. https://openai.com/blog/chatgpt

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Aguera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Karthikesalingam, A., & Natarajan, V. (2023b). Towards expert-level medical question answering with large language models.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023a). LLaMA: Open and efficient foundation language models.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hou, R., Inan, M. K., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023b). LLaMA 2: Open foundation and fine-tuned chat models.

Wu, C., Lin,W., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). PMC-LLaMA: Towards building open-source language models for medicine.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Fries, J., et al. (2022). Bigbio: A framework for data-centric biomedical natural language processing. In Advances in Neural Information Processing Systems, volume 35, pages 25792–25806. Curran Associates, Inc.

Hugging Face. (n.d.). Augmented-clinical-notes dataset. Retrieved from [https://huggingface.co/datasets/augmented-clinical-notes](https://huggingface.co/datasets/augmented-clinical-notes" \t "https://kimi.moonshot.cn/chat/_blank)

Jiang, G., et al. (2023). Mistral 7B. arXiv preprint arXiv:2301.07897.

Labrak, Y., et al. (2024). BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv preprint arXiv:2402.10373.

Singhal, K., et al. (2023a). Towards expert-level medical question answering with large language models. Nature, 620(7972), 172–180.

Touvron, H., et al. (2023b). LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2302.13971.

# 7 Appendix