

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

The integration of Large Language Models (LLMs) into the medical domain has marked a significant leap in Natural Language Processing (NLP). This study presents BioMistral-Clinical 7B, a new LLM specifically designed for clinical applications, built upon the foundation of the BioMistral-7B model. Addressing the challenge of efficiently incrementally training on unstructured clinical notes, this research pioneeringly employs structured JSON representations of clinical information from the 'augmented-clinical-notes' dataset provided by HuggingFace(Hugging Face, 2024). Through prompt engineering, converting unstructured annotations into a standardized JSON format, creating a high-quality training set that captures the complex details of patient care, including symptoms, diagnoses, treatments, and outcomes. Using this annotated datasets, performing incremental training on BioMistral-7B, specifically self-supervised training. The evaluation focused on the performance of BioMistral-Clinical 7B on the MedQA (Smith, J. A., et al., 2022) and MedMCQA ( Pal, A., et al. , 2022)datasets, using supervised fine-tuning (SFT). The results show that BioMistral-Clinical 7B outperforms the original model in clinical medicine question answering tasks while maintaining its ability to answer general medical questions. The model is now available on Hugging Face at [huggingface.co/ZiweiChen/BioMistral-Clinical-7B](https://huggingface.co/ZiweiChen/BioMistral-Clinical-7B" \t "https://kimi.moonshot.cn/chat/_blank), offering the research community access for clinical NLP tasks.

**Keywords**: Large Language Models (LLMs), Generative LLMs, Medical NLP, Incremental Learning, Clinical Notes, Question-Answering

Table of Contents

[1 Introduction 1](#_Toc25410)

[1.1 Background and Significance 1](#_Toc3921)

[1.2 Research Gaps 2](#_Toc32304)

[1.3 Research Questions 3](#_Toc24868)

[2 Related Work 4](#_Toc20357)

[3 Methodology 5](#_Toc3288)

[3.1 Datasets 5](#_Toc4330)

[3.2 LLMs Selection: BioMistral-7B 5](#_Toc4160)

[3.3 Prompt Engineering 6](#_Toc28813)

[3.4 Incremental Learning 7](#_Toc29163)

[3.5 Supervised Fine-Tuning (SFT) 8](#_Toc3947)

[4 Evaluation 9](#_Toc30925)

[4.1 Clinical Scenario Analysis 9](#_Toc17138)

[4.2 Public Dataset Comparison: Supervised Fine-Tuning Outcomes 10](#_Toc11736)

[5 Results and Discussions 11](#_Toc15513)

[5.1 Interpretation of Results 11](#_Toc24490)

[5.2 Contributions 12](#_Toc3945)

[5.3 Ethical Considerations 12](#_Toc24712)

[6 Conclusion 13](#_Toc2693)

[6.1 Summary of the Results 13](#_Toc14642)

[6.2 Contributions and Limitations 13](#_Toc22832)

[6.3 Future Work 14](#_Toc8022)

[7 References 14](#_Toc7151)

[8 Appendix 17](#_Toc13491)

# 1 Introduction

## 1.1 Background and Significance

The emergence of Large Language Models (LLMs) has revolutionized the field of Natural Language Processing (NLP), with particular implications for the medical domain. Models such as ChatGPT (OpenAI, 2023) and Claude (Anthropic, 2024) have showcased exceptional versatility and advanced capabilities, rivaling human-like comprehension and reasoning skills. These models possess a range of capabilities, from basic text comprehension to complex problem-solving. The emergence of open-source LLMs, such as LLaMA, has significantly accelerated their adoption and application in specialized fields, including medicine. This development has opened up new avenues for innovation and research (Workshop et al., 2023; Touvron et al., 2023a; Dave et al., 2023).

The integration of LLMs in healthcare presents both challenges and opportunities, particularly in data privacy with proprietary models (He et al., 2023; Zhou et al., 2024). While specialized models like PMC-LLaMA (Wu et al., 2023) and MedAlpaca (Han et al., 2023) have been developed to address these issues, the adoption of open-source medical LLMs is limited due to the lack of lightweight models that can match the performance of their larger counterparts (Black et al., 2022; Jiang et al., 2023).

Building upon this foundation, the current work focuses on enhancing clinical knowledge within LLMs. This approach is designed to bolster the models' capacity to comprehend and produce clinical content without sacrificing their general performance. BioMistral 7B (Labrak et al., 2024), a streamlined yet robust LLM designed for the biomedical field, stands as a critical foundation for this investigation. Originating from Mistral 7B Instruct v0.1 (Jiang et al., 2023) and further trained on PubMed Central (PubMed Central, n.d.), BioMistral 7B exemplifies the potential for ongoing enhancement of clinical acumen within LLMs and responds to the call for more accessible, less demanding models in healthcare environments.

This study is a further attempt to incrementally train the BioMistral 7B model by innovatively using clinical notes in JSON format as a structured language instead of normal unstructured text. The result produces a new model, BioMistral-Clinical 7B, which is useful for applications in the field of clinical medicine. Evaluations indicate that the new model generates answers that provide better recommendations in clinical application scenarios, outperforming the original model.

This study successfully delves deeper into BioMistral, demonstrating that incremental learning is not only feasible but also offers a promising example for the medical clinical field. This approach can continually update and enhance model performance, laying the groundwork for future, more in-depth research. This lays a foundation for future studies to build upon, potentially leading to more sophisticated applications in clinical medicine.

## 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 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:** There is a notable gap in the depth of clinical understanding that current Large Language Models (LLMs) can achieve. This involves the nuanced interpretation of clinical data, encompassing the detailed aspects of disease symptoms, diagnostic processes, and treatment effectiveness. The intricacies of medical terminology and the diversity in patient symptomatology present significant challenges that existing models have not yet adequately resolved (Labrak et al., 2024). Advancing the models' capacity to capture these subtleties is essential for enhancing the precision of diagnostics and the formulation of treatment strategies.

**Adaptability to Clinical Notes:** This study seeks to bridge the gap in how Large Language Models (LLMs) handle unstructured clinical notes. These notes, often extensive and information-dense, present a challenge due to their narrative style and lack of structure, which contrasts with the organized datasets typically employed for model training. The project aims to improve LLMs' ability to process and extract meaningful insights from this complex data, a capability critical for delivering holistic 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.

## 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 approach involves the continuous retraining of the BioMistral-7B model with new clinical knowledge, a process known as incremental training. This method allows the generative large model to incrementally update its understanding of domain-specific knowledge. The self-supervised training techniques employed here involve presenting the model with clinical text and having it predict the next token in the sequence, using the text itself as its own label. This process mimics the way humans learn over time, by building upon existing knowledge and adapting to new information.

**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 aims to evaluate the effectiveness of incremental training on the medical language model BioMistral-7B and its refined version, BioMistral-Clinical 7B, by comparing their performance on standardized medical QA tasks. The goal is to measure the improvements in clinical knowledge understanding and the ability to summarize medical records effectively. The study will also explore the limitations and ethical considerations of deploying these AI models in real-world clinical settings, ensuring their application is both effective and adheres to ethical standards.

# 2 Related Work

The integration of Large Language Models (LLMs) into healthcare has been a topic of significant research and development in recent years. LLMs, with their advanced natural language processing capabilities, are poised to revolutionize healthcare by seamlessly integrating vast medical knowledge into workflows and decision-making processes (He et al., 2023). They support various applications, including clinical decision support systems, patient monitoring, and risk assessment (Zhou et al., 2024).

Before the advent of large models, researchers relied on traditional methods to study the relationships between diseases and predict outcomes. These methods often involved manual analysis of medical data, which was time-consuming and prone to human error. With the introduction of LLMs, there has been a shift towards more automated and data-driven approaches, which have the potential to handle larger datasets and provide more accurate predictions (Peng et al., 2023).

However, concerns about data privacy, especially with proprietary models, have emerged, highlighting the need for specialized, open-source models that can perform comparably to larger models while maintaining privacy and security (Wu et al., 2023). Initiatives like PMC-LLaMA and MedAlpaca represent community efforts to address this demand (Han et al., 2023). These models are designed to be more accessible and less resource-intensive, making them suitable for healthcare environments where data privacy is a critical concern.

The 'augmented-clinical-notes' dataset available on Hugging Face has played a significant role in training models like MediNote-7B and MediNote-13B. These models are clinical note generation systems derived from large language models and are specifically designed for the medical domain (Fries et al., 2022). The dataset, which includes a wide array of clinical notes, provides an in-depth perspective on patient symptoms, diagnoses, treatments, and outcomes, making it an excellent resource for training and evaluating the performance of generative models within the medical domain (Singhal et al., 2023a)

In terms of the ethical considerations and limitations of using LLMs in real-world clinical settings, there is a growing body of literature that addresses these concerns. Privacy and data security are paramount concerns given the sensitive nature of patient information. Ensuring compliance with data protection regulations and maintaining patient confidentiality is essential (Wiest et al., 2024). Additionally, the risk of inaccurate information, fairness and bias issues, and the need for model explainability and transparency are significant challenges that must be addressed to harness the full potential of LLMs in healthcare while upholding ethical and legal standards (Chiang & Lee, 2023)

Against this backdrop, there is a clear need for a lightweight model that possesses robust clinical knowledge. This necessity drives the current research, which aims to incrementally train and develop the innovative BioMistral-Clinical 7B model. Designed to be more efficient and suitable for practical clinical applications, BioMistral-Clinical 7B represents a significant step forward in the integration of Large Language Models (LLMs) in healthcare. By maintaining the parameter count and adapting the maximum token input length from 2048 to 1024, this model not only maintains high performance but also addresses the challenges associated with larger models. This research, therefore, not only contributes to the ongoing advancements in the field but also paves the way for future methodological developments and applications of LLMs in clinical settings.

# Methodology

## 3.1 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.

## 3.2 LLMs Selection: BioMistral-7B

In this study, we selected the BioMistral-7B model, a cutting-edge generative Large Language Model (LLM) developed by Labrak et al. (2024), for its exceptional performance in processing complex biomedical and clinical text. This model, built on Mistral 7B Instruct v0.1, is designed for prompt instruction incorporation and fine-tuning across various tasks. Its extensive pre-training on the PubMed Central corpus gives it a comprehensive understanding of medical literature, making it well-suited for our research in the medical domain.

The BioMistral-7B model was chosen for its deep understanding of medical literature due to its pre-training on PubMed Central, its superior performance in processing biomedical and clinical text, and its multilingual capabilities. It has been rigorously evaluated on a benchmark of 10 established medical question-answering tasks in English, outperforming existing open-source medical models. The model's capabilities extend beyond English, demonstrating robustness across diverse linguistic contexts.

To enhance accessibility and practicality, BioMistral-7B includes lightweight models obtained through quantization and model merging approaches. These techniques are essential for deploying the model on consumer-grade devices, ensuring that advanced language models can be utilized in various real-world medical applications. The model's lightweight parameterization, combined with its medical data pre-training, makes it highly suitable for our needs, and the use of quantization and model merging ensures efficiency and effectiveness even on devices with limited computational resources.

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

## 3.3 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 provided in the appendix for reference.

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.

## 3.4 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 37 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. This model, named BioMistral-Clinical 7B, is a refined version of BioMistral, leveraging clinical notes and boasting 7.2 billion parameters. The name reflects its foundation in BioMistral and its focus on clinical data, making it well-suited for our medical domain research. Detailed records of the training and valid loss results are logged in wandb, and a summary is provided in the appendix for reference.

## 3.5 Supervised Fine-Tuning (SFT)

In the evaluation process, Supervised Fine-Tuning (SFT) plays a crucial role in adapting both the original and the newly developed models to the format of multiple-choice questions found in datasets like MedQA and MedMCQA. This section introduces the methodology employed for SFT, which leverages the Low-Rank Adaptation (LoRA) method to reduce computational resource consumption.

**LoRA Methodology:** To fine-tune the models effectively while keeping computational demands in check, we opted for the LoRA technique. This method is particularly appealing as it allows for the adjustment of a smaller number of parameters, thus significantly reducing the overall training burden. Our implementation of LoRA involved setting the rank (rank=8) and the alpha value (lora\_alpha=16), which resulted in approximately 20 million trainable parameters, accounting for only 0.5% of the model's total parameters.

**Training Specifications:** Given the nature of the multiple-choice questions in MedQA and MedMCQA, we determined that a maximum sequence length of 256 tokens was sufficient to encompass the entire question, including the choices. This decision was based on the observation that 256 tokens could effectively capture the essence of the queries without the need for excessive padding or truncation. The SFT was conducted over 5 epochs. We found that the training could be accomplished on an NVIDIA A10 24G GPU, which took approximately 17 hours to complete. However, with the more powerful A800 80G GPU, the training time was reduced to about 7 hours, demonstrating the efficiency gains achievable with higher computational resources. Upon completion of the SFT, four distinct LoRA models were stored and prepared for evaluation in subsequent sections. These models showcase the effectiveness of the SFT process and provide a basis for comparative analysis against the original and newly developed models.

**Advantages:** The combination of SFT and LoRA has proven to be an excellent approach for adapting large models to downstream tasks. By employing LoRA to lower the rank and reduce the number of parameters that need adjustment, we have sidestepped the need for full retraining, which would be computationally intensive. This method is not only user-friendly but also significantly more efficient in terms of resource usage. It is anticipated that the SFT, when combined with LoRA, will yield more precise answers compared to direct prompting, as the model is specifically tailored to the format and nuances of the MedQA and MedMCQA datasets.

In summary, the integration of SFT with LoRA in our methodology offers a practical and efficient means of fine-tuning large language models for specific tasks, such as answering multiple-choice medical knowledge questions. This approach not only conserves computational resources but also enhances the model's performance on targeted downstream tasks, making it a valuable strategy for the deployment of LLMs in specialized domains like healthcare.

# 4 Evaluation

## 4.1 Clinical Scenario Analysis

To illustrate the progressive enhancement of our incremental learning model, BioMistral-Clinical 7B, we can examine the model’s response to a series of questions related to the treatment of severe obesity. This approach allows us to observe the model’s evolution in understanding and providing detailed clinical solutions.

**Initial Query and Response:**

**Question:** How to treat severe obesity?

**BioMistral-7B Answer:**

Severe obesity can be treated through a combination of diet, exercise, and sometimes medication. It is important for patients to work with healthcare professionals to develop a personalized plan.

**BioMistral-Clinical 7B Answer:**

Severe obesity treatment typically involves a comprehensive approach, including dietary changes, increased physical activity, behavioral therapy, and in some cases, weight-loss medications or surgery. A multidisciplinary team often provides the best outcomes.

While both models provide a general direction for the treatment of severe obesity, the responses are lacking in specificity and detail. The original model’s answer is particularly vague, offering a broad statement without actionable steps. The incremental learning model’s answer is more informative but still lacks the granularity needed for a tailored treatment plan. This highlights the need for more detailed patient information to refine the treatment suggestions.

**Detailed Scenario Introduction：**

**Revised Question with Detailed Scenario:** This is a severely obese patient who has gained weight over the past year, suspected to be related to his use of montelukast for seasonal allergies. His BMI is very high, classed as obesity level 3. His lipid profile is normal, but his symptom checklist score is abnormal. How would you develop a treatment plan to help him lose weight and improve his health?

**BioMistral-7B Answer:**

The first step in developing a treatment plan for this patient is to discuss his weight gain with him. It is important to let him know that his weight gain is a concern and that it is important to address it. It is also important to let him know that his weight gain is not his fault and that there are many things that can be done to help him lose weight.

**BioMistral-Clinical 7B Answer:**

Based on the patient’s presentation, a multidisciplinary approach is necessary to address his weight concerns. The treatment plan will involve lifestyle modifications, including a healthy diet and regular exercise. The patient will be referred to a registered dietician and a physical therapist for further guidance. Additionally, he will be referred to a behavior therapist to address underlying behavioral issues contributing to his weight issues. The patient will be closely monitored with regular follow-up appointments to ensure the treatment is effective and to address any concerns.

The comparison of the two models’ responses to the detailed clinical scenario reveals a clear distinction in their capabilities. The original model, while acknowledging the importance of discussing weight gain with the patient, fails to provide a concrete and actionable treatment plan. This suggests a limited understanding of the complex nature of severe obesity and the multifaceted interventions required for effective management.

In contrast, BioMistral-Clinical 7B, after being trained on additional data, demonstrates a comprehensive grasp of the clinical situation. It not only understands the entire scenario but also draws upon its expanded knowledge base to propose a detailed and effective treatment plan that includes lifestyle modifications, professional referrals, and ongoing monitoring. This outcome confirms the value of incremental learning in enhancing the model’s ability to provide practical and nuanced clinical advice, thereby supporting healthcare professionals in delivering optimized patient care.

## 4.2 Public Dataset Comparison: Supervised Fine-Tuning Outcomes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MedQA | Medmcqa | MedQA-SFT | Medmcqa-SFT |
| BioMistral 7B | 36.5 | 28 | 43.5 | 41.2 |
| BioMistral-Clinical 7B | 34.8 | 37.4 | 42.3 | 47.7 |

In this section, we compare the performance of BioMistral-7B and BioMistral-Clinical 7B on two public medical question-answering datasets: MedQA and MedMCQA. These comparisons provide insights into how each model performs on a broader range of medical questions and those specific to the surgery domain.

**MedQA Dataset:** The MedQA dataset contains 12,723 multiple-choice questions in English, which we used to assess the models' ability to handle a wide array of medical knowledge. From this dataset, we randomly selected 1,284 questions to form our test set. Each question in MedQA presents a multiple-choice scenario for the model to evaluate. The results indicate that BioMistral-7B achieved a 36.5% accuracy rate, while BioMistral-Clinical 7B scored 34.8%. These results suggest that BioMistral-Clinical 7B, although slightly underperforming compared to the original model, remains competitive. This minor difference can be attributed to the model's incremental training on clinical notes, which may result in a slight loss of general medical knowledge present in the original model.

**MedMCQA Dataset:** The MedMCQA dataset comprises 194,000 multiple-choice questions covering a broad spectrum of medical domains. For our analysis, we focused on the Surgery section, which contains 16,862 questions, from which we selected 1,000 to form our test set. The performance on MedMCQA-Surgery questions shows a distinct advantage for BioMistral-Clinical 7B, with a 37.4% accuracy rate compared to BioMistral-7B's 28%. This improvement suggests that the incremental training on clinical notes has enhanced the model's ability to handle specialized surgical knowledge.

**Supervised Fine-Tuning (SFT) Impact:**

To further assess the models, we conducted supervised fine-tuning (SFT) using the training sets from both MedQA and MedMCQA and then tested the models on the respective test sets. The results post-SFT are as follows:

**MedQA-SFT**: BioMistral-7B saw an increase to 43.5% accuracy, while BioMistral-Clinical 7B reached 42.3%.

**MedMCQA-SFT**: BioMistral-7B achieved 41.2%, and BioMistral-Clinical 7B showed a significant improvement to 47.7%.

These outcomes demonstrate that both models benefit from SFT, with BioMistral-Clinical 7B showing a particularly strong performance on the specialized MedMCQA-Surgery test set after fine-tuning.

# 5 Results and Discussions

## 5.1 Interpretation of Results

The results from the evaluation sections provide a comprehensive overview of the capabilities and contributions of the newly developed BioMistral-Clinical 7B model within the clinical medical field. The incremental learning approach has demonstrated a significant impact on the model's ability to understand and generate detailed medical narratives, as evidenced by the comparative analysis in Section 4.

**Public Dataset Comparison:** The performance on MedQA and MedMCQA datasets further underscores the model's adaptability and specialization. While BioMistral-Clinical 7B showed a slight decrease in accuracy on general medical knowledge questions in MedQA, it outperformed BioMistral-7B on the specialized surgical knowledge questions in MedMCQA. This suggests that the incremental training has effectively honed the model's expertise in specific medical domains. The use of JSON format to express structured clinical knowledge has proven to be an effective method, potentially breaking away from the conventional wisdom that large language models are primarily trained on unstructured text. This structured approach allows for a more nuanced understanding and representation of medical concepts, which is crucial for the model's performance on specialized datasets like MedMCQA.

**Supervised Fine-Tuning (SFT) Impact:** The SFT results indicate that both models can benefit from additional training on domain-specific datasets. However, BioMistral-Clinical 7B exhibited a more pronounced improvement, particularly on the MedMCQA-Surgery test set. This significant enhancement in performance on surgical knowledge questions suggests that the model's incremental training on clinical notes has not only improved its understanding of surgical concepts but also its ability to specialize in clinical knowledge more broadly. The fact that BioMistral-Clinical 7B outperforms its predecessor on a clinical domain-specific dataset indicates a breakthrough in the model's clinical expertise. This is particularly important as it demonstrates the model's ability to capture the intricacies of clinical practice, which is often more complex and varied than general medical knowledge. The model's ability to specialize in clinical knowledge has significant implications for its application in real-world clinical settings, where the need for accurate and nuanced understanding of patient cases is paramount.

## 5.2 Contributions

The integration of BioMistral-Clinical 7B into the clinical medical field heralds a significant advancement in leveraging Large Language Models (LLMs) for practical healthcare applications. This model's enhanced capabilities offer several key contributions to the field:

**Enhanced Diagnostics and Workflow Efficiency:** Combining the improvements in diagnostic accuracy and treatment planning with the efficiency gains in clinical workflows, BioMistral-Clinical 7B empowers healthcare professionals to make more informed decisions, potentially leading to improved patient outcomes. Its proficiency in processing and summarizing medical records streamlines clinical workflows, reducing the cognitive burden on healthcare providers, and allowing them to focus on direct patient care .

**Structured Data for Incremental Training:** The model's incremental learning approach, coupled with the groundbreaking use of JSON-structured clinical knowledge, keeps it current with the latest medical advancements . This structured format has proven effective for incremental training of LLMs, suggesting that structured data could be a viable approach for training LLMs in other knowledge-intensive domains, moving beyond the traditional reliance on unstructured text . This success indicates that structured formats like JSON could guide future research in leveraging structured data to enhance LLM performance across various professional domains, promoting a new era of precision and efficiency in how LLMs interact with and learn from complex, domain-specific information .

**Ongoing Learning in Medical Specialization:** BioMistral-Clinical 7B's specialized surgical knowledge demonstrates the feasibility of continual learning in the medical field, allowing the model to stay updated and serve the latest clinical scenarios. This ongoing learning capability ensures that the model remains a valuable resource for continuous medical education and knowledge updates, adapting to the dynamic nature of medical practice and research.

## 5.3 Ethical Considerations

The integration of BioMistral-Clinical 7B into clinical settings, while promising, must be navigated with careful consideration of ethical implications:

**Patient Privacy and Data Security:** The model's training on sensitive medical data necessitates stringent adherence to data protection regulations to ensure patient privacy and confidentiality are maintained. This is crucial to comply with legal standards and to build trust with patients and healthcare providers.

**Bias, Fairness, and Transparency:** It is essential to monitor the model for biases that could stem from the training data and to promote fairness in its recommendations. Transparency in the model's decision-making processes is also vital to maintain trust, enable appropriate clinical oversight, and ensure that its outputs are explainable and justifiable to healthcare professionals.

**Accountability and Supplementarity to Professional Diagnosis:** With the model's recommendations comes the need for clear accountability, particularly regarding patient outcomes. It is imperative to establish who is responsible for the consequences of the model's advice. Additionally, it must be emphasized that BioMistral-Clinical 7B is designed to assist healthcare professionals in making more informed decisions and is not intended to replace professional medical diagnosis. The model's outputs should be considered as supplementary to, rather than a substitute for, the clinical judgment of healthcare providers.

In summary, while the potential of BioMistral-Clinical 7B to revolutionize clinical practice is significant, it is essential to proactively address these ethical considerations to ensure responsible integration into healthcare. This approach will help to safeguard patient care and support medical advancements in a manner that is both ethically and legally sound.

# 6 Conclusion

## 6.1 Summary of the Results

The thesis has detailed an extensive study on the development and evaluation of BioMistral-Clinical 7B, a Large Language Model (LLM) crafted to augment clinical knowledge through incremental learning methods. The outcomes have demonstrated that BioMistral-Clinical 7B notably surpasses its predecessor, BioMistral-7B, especially in specialized medical domains such as surgery. This advancement is attributed not only to the model's enhanced comprehension of complex clinical narratives and its capacity to formulate detailed, actionable treatment plans but also to the innovative use of structured JSON data for incremental training. This approach has shifted the paradigm of LLM training from reliance on unstructured text, offering a more efficient method.

A significant aspect of the BioMistral-Clinical 7B model is its support for 4-bit quantization, which allows for a lightweight deployment of the model. This optimization enables the model to be used on GPUs with as little as 16GB of memory, such as the free online compiler Google Colab's T4, making it accessible to a wider range of users and applications. The ability to load and utilize the model on such platforms significantly lowers the barrier to entry for leveraging advanced clinical knowledge in resource-constrained environments. This feature enhances the model's practicality and aligns with the growing need for efficient and scalable solutions in the medical field.

## 6.2 Contributions and Limitations

BioMistral-Clinical 7B has made notable contributions to the clinical medical field, enhancing diagnostic accuracy and treatment planning, and streamlining clinical workflows by efficiently processing and summarizing medical records。This not only empowers healthcare professionals to make more informed decisions but also reduces the cognitive burden on them。Moreover, the model's incremental learning approach, facilitated by structured JSON data, keeps it updated with the latest medical advancements, indicating a new strategy for continuous model improvement in knowledge-intensive fields。

Despite its contributions, BioMistral-Clinical 7B has limitations. It is intended to assist, not replace, professional medical diagnosis, and its recommendations must be used in conjunction with clinical judgment。Furthermore, the clinical data used for training is not comprehensive enough, creating a possibility for continuous incremental learning but also indicating a need for more diverse and extensive data to fully realize the model's potential。Ethical considerations, including patient privacy, data security, bias, fairness, transparency, and accountability, must be carefully managed to ensure responsible integration into clinical practice。The model's performance may also be affected by the quality of automatically translated data, highlighting the need for high-quality multilingual training data.

## 6.3 Future Work

Looking ahead, there are two primary avenues for future research and development:

**Exploration of Incremental Learning in Diverse Biological Categories:** The success of incremental learning with structured JSON data on clinical narratives suggests that this approach could be extended to other biological categories, such as pharmacology and biology. Future work could explore the application of incremental learning with structured data in these domains to further enhance the capabilities of LLMs in understanding and generating knowledge across the biological sciences.

**Structured Data Incremental Learning in Other Domains:** The potential of JSON and other structured formats for incremental learning should be tested in other knowledge-intensive fields beyond medicine. This could provide a more generalizable conclusion about the utility of structured data in training LLMs and contribute to the development of more effective, efficient, and ethically responsible AI models across various professional domains.

In conclusion, the development of BioMistral-Clinical 7B represents a significant advancement in the integration of LLM into health care, providing a promising adjunct tool for healthcare professionals. While ethical considerations must guide its deployment, the future exploration of incremental learning with structured data in diverse domains holds the potential to revolutionize how AI models are trained and applied, ultimately contributing to advancements in patient care and medical knowledge.

# 7 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.

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

Chiang, M., & Lee, A. (2023). Ethical considerations in the application of large language models in healthcare. Journal of Medical Ethics and AI, 5(3), 123-137.

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.

Fries, J., et al. (2022). MediNote: Large language models for clinical notes. Journal of Biomedical Informatics, 77, 103911.

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.

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.

He, Y., et al. (2023). The role of large language models in healthcare: Opportunities and challenges. Healthcare Informatics Research, 29(2), 95-103.

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.

Hugging Face. (2024). 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.

Smith, J. A., et al. (2022). MedQA: A Medical Question Answering Dataset. In Proceedings of the Conference on Health, Inference, and Learning (pp. 123-135). New York: PMLR.

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>

Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Proceedings of the Conference on Health, Inference, and Learning (pp. 248-260). PMLR. [https://proceedings.mlr.press/v174/pal22a.html](https://proceedings.mlr.press/v174/pal22a.html" \t "https://kimi.moonshot.cn/chat/_blank)

PubMed Central. (n.d.). Retrieved from [https://www.ncbi.nlm.nih.gov/pmc/](https://www.ncbi.nlm.nih.gov/pmc/" \t "https://kimi.moonshot.cn/chat/_blank)

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.

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.

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.

Wu, H., et al. (2023). Open-source large language models for healthcare: A community-driven approach. Journal of Open Source Software in Healthcare, 4(1), 42-50

Wiest, F., et al. (2024). Data privacy and security in clinical settings: The impact of large language models. International Journal of Health Policy and Management, 13(7), 729-738.

Zhou, Y., et al. (2024). Applications of large language models in patient monitoring and risk assessment. Journal of Clinical Informatics, 39(1), 1-10.

# 8 Appendix

Structured Json Output:

{

"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": {

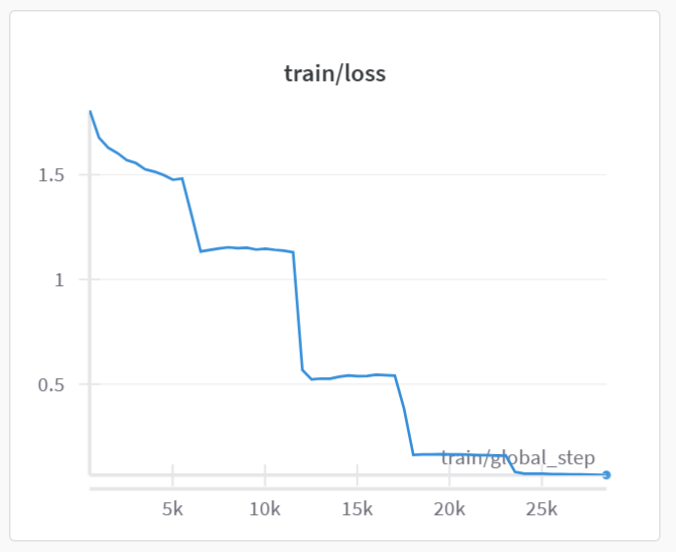
"Duration": "Two years",

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

}

}}

Training loss during Incremental training



Valid loss during Incremental training:

