

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 (PE), 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 dataset, 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), offering the research community access for clinical NLP tasks.

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

Table of Contents

[1 Introduction 1](#_Toc182780448)

[1.1 Background and Significance 1](#_Toc182780449)

[1.2 Research Gaps 1](#_Toc182780450)

[1.3 Research Questions 2](#_Toc182780451)

[2 Related Work 3](#_Toc182780452)

[3 Methodology 4](#_Toc182780453)

[3.1 Datasets 4](#_Toc182780454)

[3.2 LLMs Selection: BioMistral-7B 5](#_Toc182780455)

[3.3 Prompt Engineering 5](#_Toc182780456)

[3.4 Incremental Learning 6](#_Toc182780457)

[3.5 Supervised Fine-Tuning (SFT) 8](#_Toc182780458)

[4 Evaluation 8](#_Toc182780459)

[4.1 Clinical Scenario Analysis 8](#_Toc182780460)

[4.2 Public Dataset Comparison: Supervised Fine-Tuning Outcomes 10](#_Toc182780461)

[5 Results and Discussions 11](#_Toc182780462)

[5.1 Interpretation of Results 11](#_Toc182780463)

[5.2 Contributions 12](#_Toc182780464)

[5.3 Ethical Considerations 12](#_Toc182780465)

[6 Conclusion 13](#_Toc182780466)

[6.1 Summary of the Results 13](#_Toc182780467)

[6.2 Contributions and Limitations 13](#_Toc182780468)

[6.3 Future Work 13](#_Toc182780469)

[7 References 14](#_Toc182780470)

[8 Appendix 16](#_Toc182780471)

[9 Declaration of authorship 17](#_Toc182780472)

# 1 Introduction

## 1.1 Background and Significance

The emergence of LLMs has revolutionized the field of Natural Language Processing (NLP), with 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 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).

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 study. 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 allows for continuous updating and improvement of model performance, providing the basis for more in-depth research in the future. This builds the foundation for future research with the potential for more sophisticated applications in clinical medicine.

## 1.2 Research Gaps

The current project has the potential to fill several important research gaps in the field of medical AI, which is necessary for the development of LLM relevant to clinical expertise.

**Depth of Clinical Understanding:** There is a notable gap in the depth of clinical understanding that current 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 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 (Labrak et al., 2024) 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 is centered on developing effective prompt engineering strategies to extract key medical information from unstructured clinical notes. The objective is to identify prompts that lead to the most accurate and comprehensive data structuring. The methodology involves the use of Chain-of-Thought Prompting (Yu, Z., 2023) to explore how to effectively highlight core statements in clinical notes through prompt engineering. This enables LLMs to not only extract data but also to better understand the causal relationships within the medical domain. By understanding the mechanisms that connect symptoms, diagnoses, treatments, and outcomes, the model can process and summarize clinical narratives in a manner that aligns 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 how using a structured format like JSON can help LLMs understand new medical data better. The structured format helps the model see how different parts of medical notes are connected. This is similar to how knowledge graphs work, which are good at showing how things are related. The study will check if using this method can make LLMs understand medical data better, even when it hasn't seen that data before. This could show how using structured data can help the model learn more about medical information, especially when there aren't many knowledge graphs available.

**RQ4: How does the performance of the incremental pretrained medical LLMs compare to the original model on standardized medical question-answering tasks?**

This question aims to assess how the performance of incremental trained medical LLMs, BioMistral-7B and BioMistral-Clinical 7B on standardized medical question-answering tasks. The study will employ Supervised Fine-Tuning (SFT) on these tasks to evaluate the models' ability to understand clinical knowledge. The comparison will reveal whether the incremental training has led to improvements in performance.

# 2 Related Work

The integration of 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 (Wu et al., 2023) and MedAlpaca (Han et al., 2023) represent community efforts to address this demand. 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 HuggingFace (AGBonnet., 2024). has played a significant role in training models like MediNote-7B and MediNote-13B (Fries et al., 2022). These models are clinical note generation systems derived from LLMs and are specifically designed for the medical domain. 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 an increasingly literature that addresses these concerns. Privacy and data security are paramount concerns given the sensitive nature of patient information. It is crucial to ensure compliance with data protection regulations and maintain patient privacy (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 a new Clinical LLMs, which is BioMistral-Clinical 7B. Designed to be more efficient and suitable for practical clinical applications, BioMistral-Clinical 7B represents a step forward in the integration of LLMs in healthcare.

# Methodology

## 3.1 Datasets

The datasets employed in this study is the 'augmented-clinical-notes' datasets (AGBonnet., 2024), which is part of the HuggingFace datasets collection and can be referenced as AGBonnet/augmented-clinical-notes. This dataset 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 dataset exhibit a wide variety of medical conditions and treatments, making it an exemplary dataset for the development of a robust medical LLMs.

The length of the complete clinical notes in this dataset ranges from 746 to 31,000 words, and each note details symptoms, diagnoses, treatments, and outcomes. This dataset is not only valuable for its medical content, but also very suitable for constructing structured input for model training because the interrelationships between various medical entities can be captured. To illustrate the nature of the data, a summary of one case in the dataset is provided below, with only the essential information retained::

**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 illustrates the depth and breadth of the 'augmented-clinical-notes' dataset, highlighting its utility in training LLM to understand and generate structured medical narratives. These structured medical narratives reflect the intricacies of real-world clinical practice. The comprehensiveness of the dataset makes it an ideal resource for developing models. Therefore, the model can process and summaries clinical information in a way that is consistent with the established structure of medical knowledge.

Although this dataset is well suited due to its rich medical content, it is important to note that the notes are not structured content and the text itself is very long, averaging at 5K in length. It cannot be fed directly into the model in its entirety, which would be inefficient as it would contain a lot of irrelevant content and the model would not be able to learn much. Therefore, this study needs to make these notes more structured. That is to use the generalized LLMs Prompt Engineering (PE) like ChatGPT, which has much more powerful text summarization capabilities. By using PE, the entire dataset can be annotated and unstructured clinical notes can be transformed into structured information that can be used more effectively for model training.

## 3.2 LLMs Selection: BioMistral-7B

In this study, the BioMistral-7B (Labrak et al., 2024) model was chosen as the base model for this experiment as it performs well in processing complex biomedical and clinical texts. The model was originally built on the Mistral 7B Instruct v0.1 (Jiang et al., 2023) and was designed to quickly incorporate instructions and fine-tune various tasks. It has been extensively pre-trained on the PubMed Central corpus (PubMed Central, n.d.), which gives it a comprehensive understanding of the medical literature and is well suited to this research in the medical field.

The BioMistral-7B model was chosen as the base model for the study because it is fully pre-trained on PubMed Central, which gives it a strong understanding of biomedical literature. While the model performs well on a benchmark of 10 established English medical question-answering tasks, outperforming existing open-source medical models, it still has room for improvement in real-world clinical applications. Currently, BioMistral-7B has a broad base of general medical knowledge, making it an ideal candidate for further training to achieve greater specificity in clinical domains. This is in line with the goal of this study to improve the model for use in more specialized clinical applications, which has its existing strengths as a basis for training in more specific clinical specialty areas.

## 3.3 Prompt Engineering

The overall goal of this study is to design a structured clinical notes analysis method that improves the granularity and relational clarity of patient data. Since raw clinical notes are unstructured text, directly using them for incremental learning is inefficient and may obscure key information with irrelevant details. To address this issue, this study proposes a method to annotate these unstructured notes with the generalized LLMs and convert them into a standardized JSON format. This structured format aims to mitigate the inefficiencies associated with processing long and unformatted text by distilling information to summaries the essence of each patient case.

The approach of this study is to segment the data into different categories such as chief complaint, medical history, diagnostic findings, diagnosis, treatment and outcome, each category being delineated by a set of well-defined sub-fields. This structured JSON format will be used as a dedicated input to the model, thus effectively eliminating raw text data.

To achieve this transformation, this study adopted Prompt Engineering technology to annotate clinical notes using generalized LLMs. Because traditional LLMs are the best at text understanding, in the process of annotation, it is not necessary to understand the relationship between individual pathologies, but to have the ability to understand the connection between various parts of the text and extract the individual parts, which is exactly what it has.

A Chain-of-Thought Prompting (Yu, Z., 2023) approach to PE design was used in this study. This includes experimenting with LLMs for role definition, rule interpretation, output specification and output examples. After several iterations and comparative analysis of the performance of different models such as GPT-3.5 Turbo, GPT-4 Mini and GPT-4, all the models are effective in outputting the expected results. Based on the consideration of economic efficiency, the experiment finally selected the GPT-3.5 model for further labeling work.

The experiment finally designed the following prompt template for the clinical note’s organization task. First, the identity is given as professional clinical notes organizer, then the content and sample output of the rule are prompted in sequence, and finally the clinical notes are passed in to get the output. The content explanation and sample output are omitted in the figure:

As a professional clinical note’s organizer, my task is to take a set of clinical notes and structure them into a JSON format. This format will help in standardizing the patient data for better analysis and record-keeping. Below, I will explain each field in the JSON structure and provide an example based on a hypothetical clinical note.

JSON Field Explanation: …

Example Output: …

Medical case data:

{message}

Output Json:

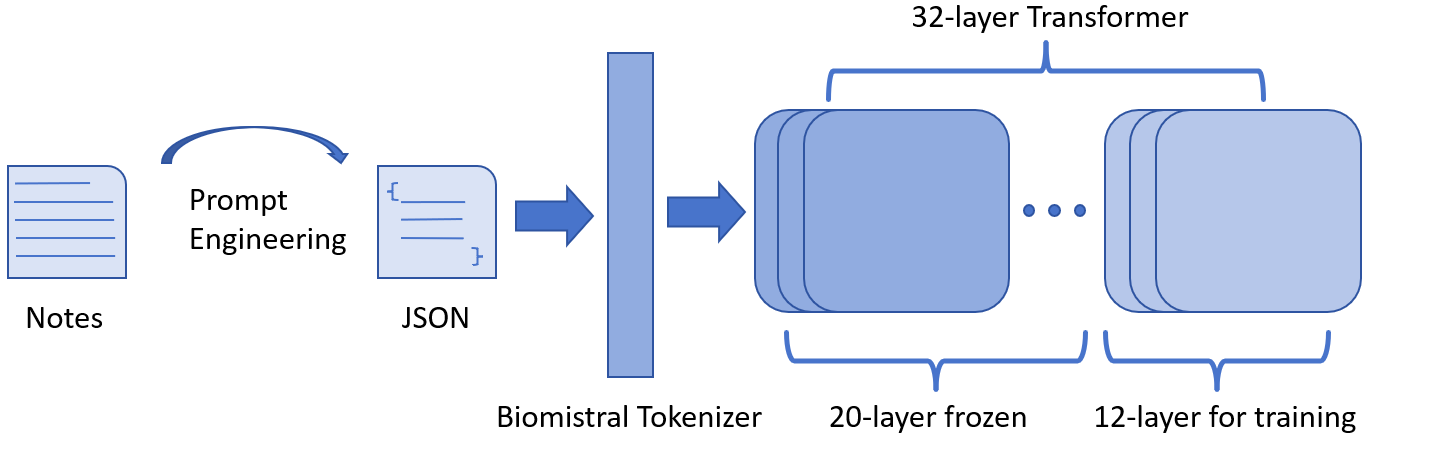
**Figure 1:** The Final Prompt templates.

In the study, the entire process of annotating 30K notes involved the use of the GPT-3.5 Turbo model, consuming a total of about 100 million tokens and taking about 40 hours to generate training data. Based on observations from manual spot checks of the results, this is a dataset that is not only well-organized, but also preserves key relationships between various pathologies, treatments, and outcomes. By providing this structured input, it will be used for incremental learning to identify patterns, correlations, and dependencies more accurately between different aspects of patient care.

Adopting this structured data format will simplify the data preprocessing stage and provide a solid foundation for building models that are scalable and adaptable to the ever-changing complexity of clinical data management. An example of structured JSON output is provided in the appendix for reference.

## 3.4 Incremental Learning

In this part of the methodology, this study focuses on the incremental learning process for the Biomistral-7B model (Labrak et al., 2024). By performing self-supervised pre-training on structured JSON data, this approach allows the model to learn implicitly without explicit labelling, to acquire new knowledge and enhance its understanding of clinical information.



**Figure 2**: The process of Incremental Learning.

**Data Preparation and Tokenization**

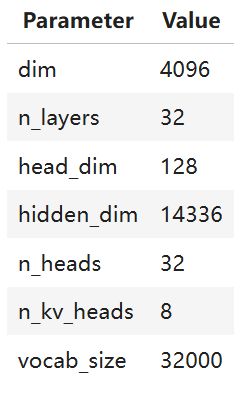
In the incremental learning stage of this study, structured JSON data obtained through the prompt engineering (PE) technique described previously was used. These data contain key medical details extracted from clinical notes and are used as the training dataset for model training.

A key aspect of data preparation was determining the appropriate sequence length for training. After analyzing the length distribution of the training data, it was decided that a maximum sequence length of 1024 bytes would be sufficient to contain all the information in the JSON formatted data. This decision was made because padding the JSON data of the training dataset to a length of 1024 bytes ensures that all relevant data is retained. Although the data came from the original medical notes, which could be as much as 30,000 words, it was compressed into a more focused and detailed JSON format after PE.

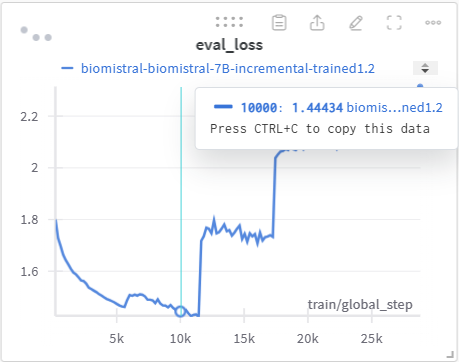
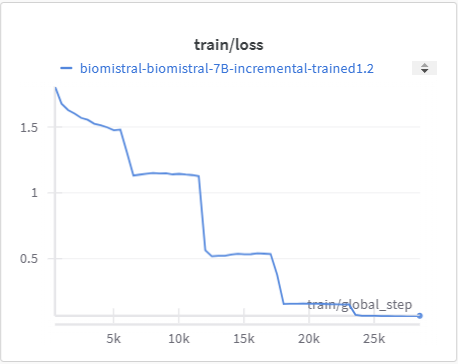
To facilitate subsequent self-supervised training of the model, the tokenizer from the original model BioMistral-7B was used before retraining. This tokenizer is used for tokenization of structured JSON data. By using the BioMistral-7B tokenizer, it ensures consistency in the tokenization process, which is essential for the model to understand and learn from the data effectively. For the training dataset is divided into training and validation sets according to 80% and 20%, the validation set is proposed to monitor the generalization ability of the model and to prevent the model from overfitting on the training set.

**Model Architecture and Training Strategy**

The model employed in this study is the BioMistral-7B, which boasts a 32-layer transformer architecture that is particularly adept at managing the medical text (Labrak et al., 2024). This deep neural network structure was chosen because of its ability to capture subtle patterns and relationships in the medical domain. To balance between maintaining the core understanding of the model and allowing for the assimilation of new information, this study chose to freeze the weights of the initial 20 layers. This strategy allowed the base layer to retain its general understanding while the top 12 layers remained unfrozen, allowing it to adapt and learn from newly introduced structured data. With only the 12 layers selectively unfrozen, the total number of trainable parameters is approximately 2 billion for the tuned 12 layers. This centralized training approach allows for more efficient use of computational resources and faster convergence during training.

After several attempts, the final training strategy was designed to have 5 epochs, which is sufficient for the model to absorb structured data and enhance its knowledge base. In terms of computational resources, the entire training process was performed on a single A800 80G GPU, which is well suited to meet the intense computational requirements of training such models. During the training process, each batch consists of 16 samples of the training dataset. As it can be seen from the backend statistics, this number of batches allows the GPU to be fully utilized, running at 100% utilization. This not only ensures the efficiency of the training, but also maximizes the throughput of the GPU, making the training process faster and more efficient. The entire training lasted 37 hours, providing enough time for the model to converge and get a full grasp of the input data.

**Results and Model Selection**



Selected

point

Point

**Figure 3:** Detailed records of training and evaluate loss results.  **Table 1：**Model architecture.

Over the approximately 30,000 steps of training, the training loss continues to decrease, indicating a strong fit as the model learns from the data. However, the effective loss shows a different trend: it initially decreases, reaches a minimum at about 11,000 steps, and then rises again, suggesting that overfitting is beginning to occur. In response to this observation, it was decided to use the 10,000-step checkpoint as the final model, as it represented the best performance point on the validation set. This decision was made to avoid overfitting and to ensure that the model generalized well to unseen data. The selected checkpoint thus becomes a trained model that can learn from new data while maintaining the original capabilities of the model. The model, named BioMistral-Clinical 7B, is a re-refined version of BioMistral that utilizes clinical notes and has 7.2 billion parameters. The name reflects the fact that it is based on BioMistral with a focus on clinical data and is well suited to current research in the medical field. Detailed records of training and evaluate loss results are documented in wandb, and a summary is provided above.

## 3.5 Supervised Fine-Tuning (SFT)

After obtaining a new model, essential is the evaluation phase. During the evaluation process, supervised fine-tuning (SFT) plays a crucial role in adapting both the original and the newly developed models to the public multiple-choice format. This section describes the methodology employed for SFT, which also utilizes the Low-Rank Adaptation (LoRA) (Hu, E. J., & Li, D., 2021) approach to reduce computational resource consumption.

**LoRA Methodology:** To efficiently fine-tune the model while controlling the computational demands, the LoRA technique was chosen for this experiment. This approach is particularly appealing because it allows for the tuning of fewer parameters, thus greatly reducing the overall training burden. The implementation of LoRA in the experiment consisted of setting the rank (rank=8) and alpha value (lora\_alpha=16), which resulted in the generation of approximately 20 million trainable parameters, which represents only 0.5% of the total model parameters.

**Training Specifications:** After analyzing the length of the MedQA and MedMCQA datasets, the experiments were decided to use a maximum sequence length of 256 tokens, which is sufficient to contain the entire QA. SFT was performed over 5 epochs, and the experiments found that it took approximately 17 hours to complete the training using the NVIDIA A10 24G GPU. However, with the more powerful A800 80G GPU, the training time was reduced to approximately 7 hours, demonstrating the efficiency gains that can be made by using higher computational resources. After completing the SFT, four different LoRA models were stored ready to be evaluated in subsequent sections. These models demonstrate the effectiveness of the SFT process and provide a basis for comparative analyses of the original and newly developed models.

**Advantages:** In fact, combining SFT and LoRA is an excellent way to adapt LLMs to downstream tasks. By using LoRA to lower the rank and reduce the number of parameters that need to be adapted, training avoids the need for computationally intensive full retraining. This approach is not only user-friendly, but also significantly more efficient in terms of resource usage. With the resulting fine-tuned model, more accurate answers will be obtained than with direct prompting, as the model is specifically tailored to the format of the MedQA and MedMCQA datasets.

# 4 Evaluation

## 4.1 Clinical Scenario Analysis

This section is designed to evaluate whether the incremental learning model, BioMistral-Clinical 7B (Labrak et al., 2024), has been enhanced. Specifically, this experiment was designed in order to observe the model's response to a range of problems associated with the treatment of severe obesity, which is a very common clinical case.

**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.

**Figure 4**: The initial query and response of two models.

While both models provide a general direction for the treatment of severe obesity, the answers provided lack specificity and detail. The BioMistral-7B’s answer was particularly vague, providing only a general statement with no actionable steps. The BioMistral-Clinical 7B 's answers were much more informative, but still lacked the detail needed to customize a treatment plan. This highlights the need for more detailed patient information to improve treatment recommendations.

**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.

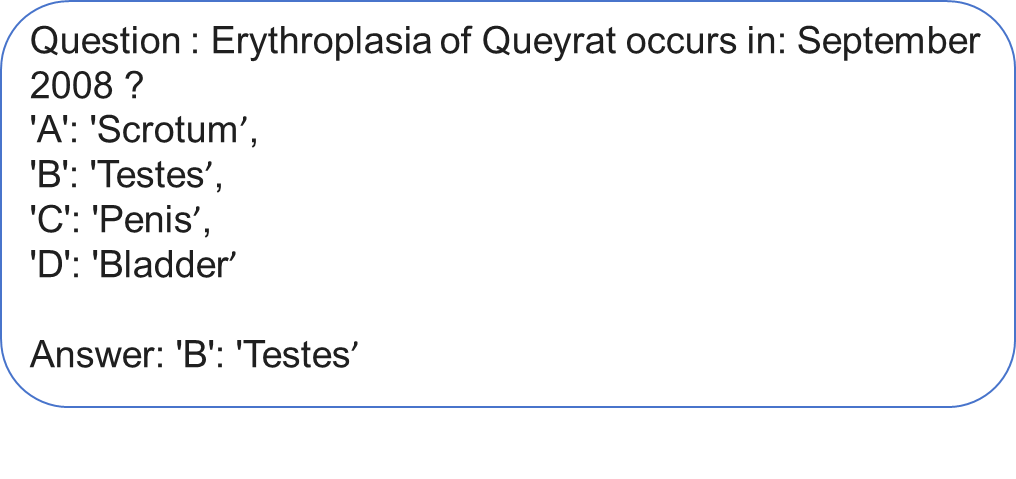
**Figure 4**: The Revised query and response of two models.

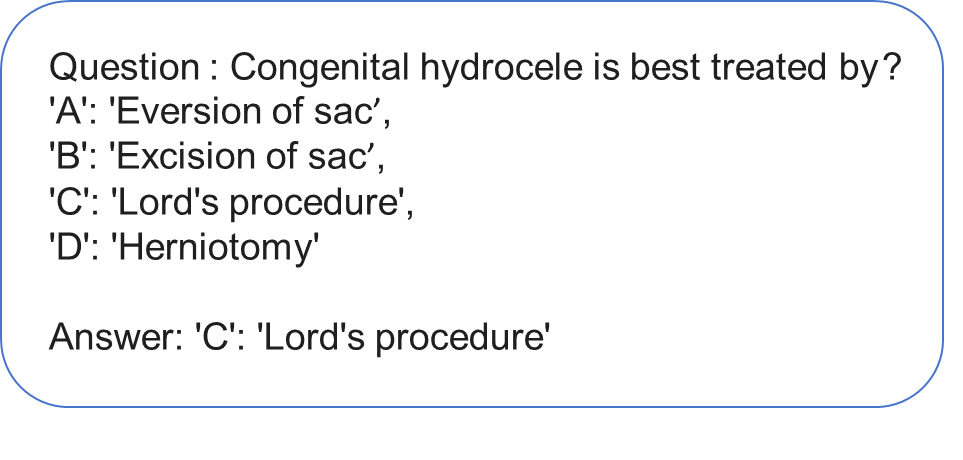
By comparing the responses of the two models to detailed clinical scenarios, there is an obvious difference in their abilities. The original model, despite acknowledging the importance of discussing the problem of weight gain with the patient, fails to provide specific and feasible treatment options. This demonstrates the complexity of severe obesity and the variability of real clinical situations.

In contrast, the BioMistral-Clinical 7B, with additional data training, can take a more comprehensive view of the clinical situation. It not only understands the entire scenario, but also proposes detailed and effective treatment options using an extended knowledge base. Specifically, this included lifestyle modifications, professional referrals, and ongoing monitoring. The comparison of this scenario demonstrates the value of incremental learning in improving the model's ability to provide practical, detailed clinical advice.

## 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** |

**Table 2:** The results of Public Dataset Comparison.



**Figure 6**: The example question of MedQA **Figure 7:** The example question of MedMCQA

In this section, this study will evaluate BioMistral-Clinical 7B. Specifically, the performance of BioMistral-7B (Labrak et al., 2024) and BioMistral-Clinical 7B is compared on two public medical question-answering datasets: the MedQA and the MedMCQA. These comparisons provide insights into the performance of each model on a broader range of medical questions and surgical domain-specific questions.

**MedQA Dataset (Smith, J. A., et al., 2022):** The MedQA dataset contains 12,723 multiple-choice questions in English, which will be used in this study to assess the model's ability to process a wide range of medical knowledge. Each question in MedQA provides a multiple-choice scenario for the model to evaluate. A test set of 1,284 questions was first randomly selected from this dataset. They were then sequentially fed to the model to generate answer choices. The results showed that BioMistral-7B had an accuracy of 36.5%, while BioMistral-Clinical 7B had an accuracy of 34.8%. From the results, it can be concluded that BioMistral-Clinical 7B is still competitive, although it is slightly inferior to the original model. This slight difference can be attributed to the increasing training of the model on clinical notes, which may result in a slight loss of general medical knowledge present in the original model.

**MedMCQA Dataset (Pal, A., et al., 2022):** The MedMCQA dataset consists of 194,000 multiple-choice questions covering a wide range of medical domains. For the study, only the surgical portion of the data was selected as the model was incrementally trained on clinical data. This section contains 16,862 questions named MedMCQA-Surgery from which 1,000 questions were selected by the experiment to form the test set. The same methodology was used to generate the answers, and as a result BioMistral-Clinical 7B's performance on the MedMCQA-Surgery questions showed a clear advantage, with an accuracy of 37.4% compared to 28% for BioMistral-7B. This improvement suggests that continuous training on clinical notes enhances the model's ability to process specialist surgical knowledge.

**After Supervised Fine-Tuning:**

To further evaluate the models, this study performed supervised fine-tuning (SFT) on each of the two models using the training sets of MedQA and MedMCQA, and then tested the models on the corresponding test sets. The results of the fine-tuning are shown below:

**MedQA-SFT**: The accuracy of BioMistral-7B improved to 43.5%, while BioMistral-Clinical 7B reached 42.3%. After SFT, the gap between the accuracy of the two models for the MedQA dataset is decreasing. It shows that BioMistral-Clinical 7B has not lost the ability of medical knowledge of the original model.

**MedMCQA-SFT**: BioMistral-7B reached 41.2 %, while BioMistral-Clinical 7B increased significantly to 47.7 %. From this result, BioMistral-Clinical 7B indeed demonstrates a stronger ability in clinical aspects after incremental learning of clinical knowledge.

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

This section will summaries the results of the entire evaluation and will discuss the contributions made by this study, and the possible ethical considerations involved in the model.

## 5.1 Interpretation of Results

**Public Dataset Comparison:** Performance on the MedQA (Smith, J. A., et al., 2022) and MedMCQA (Pal, A., et al., 2022) datasets further highlights the adaptability and specialization of the model. While BioMistral-Clinical 7B showed a slight decrease in accuracy on general medical knowledge questions in MedQA, it outperformed BioMistral-7B on specialized surgical knowledge questions in MedMCQA. This indicates that incremental training effectively sharpens the model's expertise in specific medical domains. The use of JSON format to express structured clinical knowledge proved to be an effective approach and a new attempt. This structured approach allows for a more relevant representation between medical concepts, which is crucial for model performance on public assessment datasets.

**Supervised Fine-Tuning (SFT) Impact:** The SFT results show that additional training on domain-specific datasets can benefit both models. However, the improvement of BioMistral-Clinical 7B was more noticeable, especially on the MedMCQA-Surgery test set. The significant improvement on the surgical knowledge questions indicates that incremental training of the model on clinical notes improved not only its understanding of surgical concepts, but also its ability to specialize in broader clinical knowledge. According to this result, the model's ability to capture intricate clinical practice was demonstrated. Clinical practice is often more complex and variable than general medical knowledge. Therefore, the ability to specialize in clinical knowledge is important for its application in real clinical situations.

## 5.2 Contributions

The integration of BioMistral-Clinical 7B with the field of clinical medicine is a new attempt to use LLM for practical medical applications. This research has made several potential contributions to the field:

**Enhanced Diagnostics and Workflow Efficiency:** BioMistral-Clinical 7B has the capability to improve diagnostic accuracy and treatment planning. It enables healthcare professionals to make more knowledgeable decisions that improve patient care outcomes. It simplifies clinical workflow in processing and summarizing medical records. Meanwhile, it reduces the cognitive workload of healthcare providers, allowing them to focus on direct patient care.

**Structured Data for Incremental Training:** In this experiment, an experimental use of JSON structured clinical knowledge was made in the incremental training of the model. This structured format proved to be very effective for incremental training of the LLM, suggesting that structured data can be a viable method for training LLMs in other knowledge-intensive domains, thus moving beyond the reliance on traditional text. This success suggests that structured formats such as JSON can guide future research to try this approach to improve LLM performance in various specialized areas.

**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

While the integration of BioMistral-Clinical 7B into the clinical setting is promising, the ethical implications must be carefully considered:

**Patient Privacy and Data Security:** The application of this model to sensitive medical data must strictly comply with data protection regulations to ensure that patient privacy and confidentiality are maintained. This is essential to comply with legal standards and build trust with patients and healthcare providers.

**Bias, Fairness, and Transparency:** A key aspect of this study is to be vigilant in evaluating potential biases in the model that may originate from the training dataset. In addition, transparency in modeling decisions is critical for several reasons: it fosters trust, promotes clinical oversight, and ensures that model results are both interpretable and defensible to healthcare practitioners. Emphasizing model accountability is critical to the ethical deployment of AI in healthcare settings.

**Accountability and Supplementarily to Professional Diagnosis:** With modeling recommendations comes the need for clear responsibilities, especially regarding patient outcomes. It is important to identify who should be held responsible for the consequences of the model's recommendations. In addition, it is essential to emphasize that the purpose of BioMistral-Clinical 7B is to assist healthcare professionals in making more informed decisions, not to replace professional medical diagnosis. The outputs of the model should be viewed as a supplement to, not a substitute for, the clinical judgment of healthcare professionals.

# 6 Conclusion

## 6.1 Summary of the Results

This thesis details research into the development and evaluation of BioMistral-Clinical 7B, a new LLM based on BioMistral-7B (Labrak et al., 2024) enhanced with clinical knowledge through incremental learning approach. The results of the study show that BioMistral-Clinical 7B surpasses its base model, BioMistral-7B, in terms of medical and especially clinical capabilities. The capabilities are reflected in the model's enhanced ability to understand complex clinical narratives and to develop detailed, actionable treatment plans. This enhancement is attributed to the innovative use of structured JSON data for incremental training rather than long and impure text. This approach demonstrates a new training schema for LLMs that no longer relies on unstructured text, providing a more efficient method.

An important aspect of the BioMistral-Clinical 7B model is the support for 4-bit quantization, which enables lightweight deployment of the model. This optimization allows the model to be used on GPUs with as little as 16GB of memory (e.g., the free online compiler Google Colab's T4), making the model available to a wider range of users and applications. The ability to load and use the model on such platforms significantly lowers the barrier to utilizing advanced clinical knowledge in resource-limited environments. This capability enhances the utility of the model and meets the growing demand for efficient, scalable solutions in the healthcare field.

## 6.2 Contributions and Limitations

BioMistral-Clinical 7B enhances clinical medicine by refining diagnostic precision and treatment strategies. It streamlines clinical processes through the efficient management of medical records and keeps the model current with the latest medical findings using structured data. This incremental learning approach offers a novel strategy for the ongoing enhancement of models in data-rich fields.

Despite the contributions of BioMistral-Clinical 7B, it also has limitations. It is intended to assist rather than replace professional medical diagnosis, and its recommendations must be used in conjunction with clinical judgment. In addition, the clinical data used for training is less than comprehensive, which creates the possibility for ongoing incremental learning. Ethical considerations including patient privacy, data security, bias, fairness, transparency, and accountability must be carefully managed to ensure responsible integration into clinical practice. Model 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:

**Exploring Incremental Learning in Different Medical Categories:** The success of incremental learning from clinical treatment records using structured JSON data suggests that this approach can be extended to other medical categories such as pharmacology and biology. Future work could explore the use of structured data for incremental learning in these domains to further improve LLM's ability to understand and generate knowledge in the biological sciences. For example, using JSON to represent pharmacology knowledge graphs allows for specialized model training that can be used to predict new drugs. By structuring drug data within a JSON framework, models can be trained to understand the complex relationships between drugs, targets, and diseases to discover new treatments.

**Structured Data Incremental Learning in Other Domains:** The potential of JSON and other structured formats for incremental learning should indeed be tested in other knowledge-intensive fields outside of medicine. More of these attempts could lead to more general conclusions about the utility of structured data in training LLMs. For example, in the financial field structured data formats have been used to manage transaction data. Because of their flexible structure and high performance, large amounts of financial data can be efficiently processed and analyzed. In addition, fields such as law and legal theory can benefit from LLMs trained on structured data, as they can draft contracts, analyze judgments, and assist in case studies. By expanding the application of structured data in these fields, the accessibility and usefulness of LLMs can be enhanced. Future research will prove this to be a powerful tool not only in medicine, but also in finance, law, and other fields.

In summary, the development of BioMistral-Clinical 7B represents an important attempt to integrate LLM into the healthcare field, providing a promising auxiliary tool for healthcare. Although its deployment must be guided by ethical considerations, future explorations of incremental learning using structured data in different domains have the potential to revolutionize the way AI models are trained and applied.

# 7 References

AGBonnet. (2024). Augmented Clinical Notes dataset [Dataset]. Hugging Face. <https://hf-mirror.com/datasets/AGBonnet/augmented-clinical-notes>

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.

Hu, E. J., & Li, D. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685. https://arxiv.org/abs/2106.09685

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

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>

Labrak, Y., et al. (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>

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>

PubMed Central. (n.d.). Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/>

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.

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.

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.

Yu, Z. (2023). Towards Better Chain-of-Thought Prompting Strategies: A Survey. arXiv preprint arXiv:2310.04959. Retrieved from https://arxiv.org/abs/2310.04959

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"

}

}}

# 9 Declaration of authorship

“I hereby declare

- that I have written this writing sample (thesis, seminar paper or other) without any

help from others and without the use of documents and aids other than those

stated in the references,

- that I have mentioned all the sources used and that I have cited them correctly

according to established academic citation rules,

- that the topic or parts of it are not already the object of any work or examination of

another course unless this is explicitly stated,

- that I am aware that my work can be electronically checked for plagiarism and that

I hereby grant the University of St.Gallen copyright as far as this is required for this

administrative action.”

Date and Signature

………………………………………………